Modern psychometric perspectives on the evaluation of clinical scales

Brouwer, Danny

Modern Psychometric Perspectives on the Evaluation of Clinical Scales

Danny Brouwer

ISBN: 978-90-367-6117-8
ISBN electronic version: 978-90-367-6116-1
Cover designed by We Are Bob, Bob Jansen
Printed by Ipskamp Drukkers B.V., Enschede

RIJKSUNIVERSITEIT GRONINGEN

Modern Psychometric Perspectives on the Evaluation of Clinical Scales

Proefschrift

ter verkrijging van het doctoraat in de
Gedrags- en Maatschappijwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
donderdag 11 april 2013
om 14.30 uur

door

Danny Brouwer

geboren op 21 juli 1983
te Haarlemmermeer

Promotor:	Prof. dr. R.R. Meijer
Copromotor:	Dr. J. Zevalkink
Beoordelingscommissie:	Prof. dr. P. de Jonge
	Prof. dr. J.H. Kamphuis
	Prof. dr. K. Sijtsma

Contents

Chapter 1
Introduction

1.1 Screening Procedures in Clinical Practice

Clinical psychologists use psychological tests and questionnaires at various stages of the clinical diagnostic and treatment process. In this thesis I focus on self-report questionnaires that are used as screening instruments, for example, as part of the first screening procedure for patients who seek help for their psychological problems. The purpose of a screening procedure in many mental health clinics is (1) to conduct an initial diagnosis about the psychopathology and strengths and weaknesses in the personality structure and living environment of patients, (2) to conclude whether there is a potential treatment success, and, if so (3) to consider whether to continue with the diagnostic process or to start with a particular type of treatment. If the psychologist seems unable to help, patients are referred to other clinics or psychologists. A usual screening procedure consists of a semi-structured interview conducted by a psychologist and a battery of self-report diagnostic assessment questionnaires. In a consensus meeting with several other specialists from different disciplines hypotheses about the psychopathology, personality, environment, and their interactions are evaluated. The outcomes of these meetings are further discussed with the patient and subsequent actions are taken.

To generate hypotheses and make well-informed decisions, clinical psychologists usually combine observations and information from semi-structured interviews with the scores of diagnostic assessment questionnaires. For example, consider a 22-year old student I met in my profession as a clinical psychologist (several characteristics of this case were changed to guarantee anonymity). During the first interview she made a friendly, modest, dependent, and tense impression. She explained that several years ago she had been in a successful treatment for an anxiety disorder, but that during the last months some of the symptoms were coming back. She felt restless, slightly in panic, and the compulsive rumination about her study was increasing her anxiety. She stated that she needed immediate help because the anxiety feelings were becoming more intense, but that she also would like to get to the roots of her

problems. She realized that the anxiety and rumination had affected many areas in her life for a long time. She described a highly protective youth and presented her parents as tensed and often worried. The Symptom Checklist-90 (Arrindell & Ettema, 1986; Derogatis, 1983) showed a score pattern that was similar to score patterns of a normal population of nonclinical patients, only the score on the feelings of inferiority subscale was above average. Figure 1.1 shows that the Beck Depression Inventory-II (BDI-II; Beck, Steer & Brown, 1996) depression scores were minimal and Figure 1.2 shows that the Inventory of Interpersonal Problems 64 (IIP-64; Horowitz, Alden, Wiggins, & Pincus, 2000) did not indicate severe interpersonal problems, but that the subscale scores for problems due to a non-assertive and exploitable interpersonal attitude were above average.

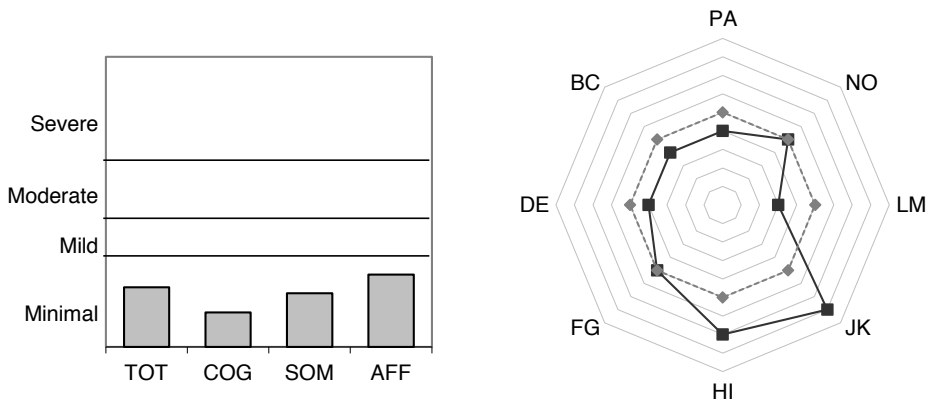

Figure 1.1: Beck Depression Inventory II and Inventory of Interpersonal Problems 64 scores The BDI-II total score and subscores are presented on the left side. TOT = Total scale, COG = Cognitive, SOM = Somatic, AFF = Affective subscale. On the right side the web of rings represents IIP-64 ipsatized stanine values on these subscales ranging from 1 for the inner ring (very low) to 9 for the outer ring (very high on problems). The grey dotted line represents the average value for a clinical population. The dark dots represent subscale values for PA = Domineering, BC = Vindictive, DE = Cold, FG = Socially Avoidant, HI = Nonassertive, JK = Exploitable, LM = Overly Nurturant, NO = Intrusive.

The combination of observations and information from the interview together with the scores on these questionnaires lead to a number of hypotheses and decisions. For example, one hypothesis was that the anxiety throughout her life could be explained in terms of a lack of confidence in her own capacities and an overly dependent perception of herself in relation to others. In psychodynamic terms this pointed to a stagnated separation-individuation process

due to an identity development characterized by overly protective parenting and a lack of encouragement to explore the world around her. This hypothesis was based on the over-friendly and dependent behavior during the interview and the information about current and past relationships. The scores of the IIP-64 indicated interpersonal distress due to a dependent and submissive attitude, which strengthened the hypothesis. In concordance with the wishes of the patient we decided to plan a consultation with a psychiatrist to discuss the use of medication to reduce the immediate anxiety. Furthermore, we decided to continue with the next step in the diagnostic process that consisted of a more thorough investigation of her personality, to determine whether she could benefit and was motivated for a long-term psychoanalytic individual or group treatment. One of the treatment goals would be to facilitate the separation-individuation process.

Although different sources of information were used for making these decisions, scores from clinical scales and subscales were considered in the diagnostic and decision-making process. In this thesis I describe how the results from modern psychometric methods can create new perspectives on the quality of psychological assessment scales that are used to substantiate clinical decision making, such as for this patient. The main aim of this thesis is to investigate how research results based on modern psychometric methods can be used to advance our thinking about assessment in clinical psychology. In this introduction I discuss three main and overarching issues in this field.

First, the use of subscales in screening questionnaires is discussed. Mental health specialists use the scores on subscales because they want to explore specific areas of psychological functioning of their patients in more detail. However, subscale scores are often based on a small number of items, which may result in unreliable measurement and furthermore there may be a large common factor underlying different subscale scores. That is, it is sometimes unclear to what degree the content of one subscale is different from the content of other subscales. Second, I discuss the factor structure of clinical scales. Clinical scales are often constructed to represent particular psychological problem areas that are heterogeneous in content. For example, a clinical scale that is aimed at measuring depression, consists of items that measure different symptoms of depression, such as sadness, agitation, changes in sleep, self-dislike, and tiredness. Thus, clinical scales are generally not strictly unidimensional and this heterogeneous nature of many clinical scales can be investigated in various ways. Third,

in many clinical research articles and test manuals it is assumed that measurement precision of individual scale scores is the same across persons. For example, often the standard error of measurement is used based on an estimated Cronbach's alpha. Several studies showed that the standard error differs across scale scores (e.g., Cole et al., 2011; Reise & Haviland, 2005; Reise & Waller, 2009). Using modern psychometric tools it is possible to provide more detailed information about measurement precision at the individual level. I discuss the advantages of these new approaches for the measurement of change. In this thesis I apply different psychometric methods to analyze clinical data.

In a humorous presidential address to the psychometric society Cronbach (1954) also reported on a Psychometric mission to Clinicia. He described two planets inhabited by two different races: Psychometrika was inhabited by the Psychometrikans and Clinicia by the Clinicians. Unfortunately, these races rarely interacted. Cronbach (1954, p. 266) said that "*Psychometrikans are in a peculiarly good position to help the Clinician just because they take quite different views of the world. The Psychometrikan views the world as one of simple relationships. Wherever he looks he perceives linear regressions, unit weights, and orthogonal variables. On the other hand the Clinician's first premise is that nature is complicated, too complicated to be caught in a simple net. No scientific generalization can take enough things into account to satisfy the Clinician. Neither philosophy is more correct than the other. The Clinician's passion for complexity is almost certainly a valid way to conceive of the universe. The Psychometrikan's passion for reduction is a practical compromise, to simplify problems enough so that scientific methods can come to grips with them*". In a more recent presidential address, Sijtsma (2012) again reminded us of the fruitful potential of a more extensive collaboration between psychometricians and clinical psychologists.

Thus far, the relative contribution of modern psychometrics, as compared to classical approaches, in the practical field of clinical psychology is modest. Throughout this thesis, I discuss and demonstrate how recent developments in psychometrics can be used to assess the quality of psychological screening instruments in a clinical useful way.

1.2 The Use of Subscales in Clinical Psychology

Psychologists combine information from observations, interviews, and results from clinical self-report questionnaires to obtain an impression of a particular patient. Results from questionnaires are based on total scores, subscale scores, and sometimes item scores. Questionnaires often consist of subscales with a varying number of items that cover specific content areas of the overall construct that is being measured. Often subscale scores are based on the responses to four to eight items. Being both a researcher and a clinician, I was curious to know to what degree the total scores and subscale scores measured different things. Can we use subscales scores to reliably differentiate between specific content areas? In clinical practice for example, the BDI-II provides a total score on Depression and separate scores on subscales such as Cognitive, Affective, and Somatic aspects of depression symptoms. Patients often ask me how to interpret the differences between these subscale scores. Also, when we discuss these subscale scores among therapists it is often unclear how to interpret subscale scores in relation to each other or in relation to the total test score.

1.3 Factor Structure of Clinical Scales

In the clinical assessment literature there are many studies that make different recommendations with result to the use of subscale scores and the underlying factor structure for specific questionnaires. For example, for the widely used BDI-II different factor structures have been proposed and the different findings have been defended through, sometimes, completely opposite arguments (e.g., Vanheule, Desmet, Groenvynck, Rosseel, & Fontaine, 2008; Quilty, Zhang, & Bagby, 2010). Another example can be found in the research literature about the factor structure of the Dispositional Hope Scale (DHS; Snyder et al., 1991). Some researchers favor the use of different subscales and encourage researchers to further study the usefulness of these subscales (Babyak, Snyder, & Yoshinobu, 1993; Chang, 2003; Snyder et al., 1991), whereas others conclude that a distinction between these subscales cannot be defended and that research that correlate the scores on these separate subscales to external criteria should be abandoned (Roesch & Vaughn, 2006; Arnau, Rosen, Finch, Rhudy, & Fortunato, 2007). From these factor analytic studies and my clinical experience I concluded that it is unclear for patients, clinical psychologists, and researchers alike, how to interpret scores based on subscales as distinct constructs of psychological functioning.

Many psychological constructs operate at different levels of generality ranging from broadband constructs to conceptually narrow constructs (Brunner, Nagy, & Wilhelm, 2012). This is inherent to the nature of complex constructs in an area such as clinical psychology. Researchers try to create scales that capture the complexity of a psychological construct and at the same time they suggest that subscales can be used to reflect more homogeneous constructs. For example, to capture depression researchers need to include items that represent the different symptoms (such as changes in sleeping pattern, eating behaviour, and suicidal thoughts) that are part of a depression syndrome. These items share content that is common to all depression items, but at the same time there are subgroups of items that are more related to each other as compared to the other items of the depression scale. Consequently, the factor structures of these scales are often not clearly one- or multidimensional. This observation may partly explain the sometimes opposite findings in the literature.

1.4 Measurement of Individual Change in Clinical Psychology

In the last decade, there is an increasing interest of clinical psychologists and policy-makers to routinely track therapy progress and outcome (e.g., de Beurs, 2012; Lambert, 2007; Lambert et al., 2003; Percevic, Lambert, & Kordy, 2006). Clinical assessment tools originally developed for screening purposes are now being used to track changes in psychological functioning during treatment. Clinical psychologists and policy-makers need results from empirical research in order to understand and to interpret the conditions under which the scores on these screening instruments are reliable and valid indicators of (change in) psychological functioning. These conditions can be investigated in different ways. For example, measurement accuracy for different scale scores can be reported through one reliability coefficient and one standard error of measurement. However, various studies have shown that measurement accuracy may differ across scale scores (e.g., Cole et al., 2011; Reise & Haviland, 2005; Reise & Waller, 2009). Modern psychometric methods are available that allow researchers to report different standard errors for different scores rather than use the same standard error of measurement.

1.5 Modern Psychometrics

Psychometrics is concerned with modeling response behavior to psychological and educational tests and questionnaires and plays an important role in the development and evaluation of many clinical instruments. As I discussed above, the domains of interest in clinical psychology are complex and most often difficult to measure. Psychometricians can help clinicians to properly define and quantify psychological constructs. From the first attempts to empirically measure psychological constructs, such as the measurement of intelligence, the dominant approach to constructing psychological tests is based on Classical Test Theory (CTT; e.g., Lord & Novick, 1968; Novick, 1966). Central in CTT is the idea that the observed scale score is the result of a true score and a random error component. The true score is defined as the average score an individual would receive when administered the same test with brainwashing in between. An alternative approach to CTT that is becoming increasingly popular among practitioners is latent trait theory, also known as item response theory (IRT; Embretson & Reise, 2000).

Although there are similarities, the main idea of IRT is fundamentally different from CTT (for more extensive descriptions and comparisons of these theories, see Embretson & Reise, 2000; Hayes, Morales, & Reise, 2000; Reise & Henson, 2003; Thomas, 2011a, 2011b). In IRT, the main assumption is that a latent variable, such as depression, cannot be directly observed but that through the observed responses to the items of a test knowledge about a person's position on the latent variable can be inferred. In an IRT model the relationship is described between a person's position on the latent variable (a person characteristic) and the probability of a response to an item, given the specific item characteristics. There are different IRT models that vary in complexity, such as the one-, two-, and three-parameter logistic model for items with dichotomous items, and the nominal response model, the partial credit model, and the rating scale model for polytomous items (for a description of these models see Embretson & Reise, 2000; Hays, Morales, & Reise, 2000). Some of these models can be extended to include multidimensional data (e.g., Cai, Yang, & Hansen, 2011; Kelderman, 1997; Reckase, 1997).

In this thesis, two psychometric models, the Graded Response Model (GRM; Samejima, 1969) and the bifactor model (Holzinger & Swineford, 1937) are of special interest because

they can be used to address the reliability and dimensionality issues discussed above. Although both psychometric models are not new, in the last decade there is an increasing interest to apply these models to describe personality and psychopathology data (Embretson & Reise, 2000; Emons, Meijer, & Denollet, 2007; Gustafsson & Åberg-Bengtsson, 2010; Reise, 2012; Reise, Morizot, & Hays, 2007; Reise, Moore & Haviland, 2010; Thomas, 2011a, 2011b).

1.5.1 The Graded Response Model

In clinical psychology many questionnaires consists of items with Likert type response categories. Likert scale items have more than two ordered response categories (for example ranging from 1 ‘strongly disagree’ through 5 ‘strongly agree’) and the answers to each category can be modeled by the GRM (e.g., Cole et al., 2011; Emons et al., 2007; Purpura et al., 2010; Walters, Hagman, & Cohn, 2011; Wu, King, Witkiewitz, Racz, & McMahon, 2012). The GRM is also used in two of the studies in this thesis. The GRM defines items by a slope parameter and two or more location parameters. The magnitude of the slope parameter reflects the degree to which the item is related to the underlying latent variable. This means that for high values the response categories accurately differentiate among latent variable levels. The location parameters reflect the spacing of the response categories along the latent variable scale. The location parameter for each category can be interpreted as the point at the latent scale where there is a 50% change of scoring that category or higher. These parameters can be used to determine the relation between the probability of a response in a particular response category conditional on the latent variable. This is the category response functions (CRF). Figure 1.2 shows the CRF’s for Item 7 ‘Introduce myself to new people’ of the socially avoidant (FG) scale of the IIP-64. Moving from the lower to the higher end of the latent variable scale shows that responding in the 0-category (no problem to introduce myself to new people) is most likely for persons who score below average on this scale. When persons become more socially avoidant the probability of responding in higher categories increases, first the 1-category, followed by the 2- and 3-category, and finally the 4-category for persons with the highest level of avoidance. The Item Information Curve (IIC) on the right side of Figure 1.2 shows that the spread of the item information and the location on the trait continuum where information is peaked are determined by the between-category threshold parameters. Generally speaking, items with higher slope parameters provide more

item information and the location parameters determine where the information is located. Item information can be added across items to form a scale information curve that is inversely related to the standard error of scale scores. This illuminates an important difference between CTT and IRT. In CTT the standard error of measurement is equal across all score levels, whereas in IRT the standard error is dependent on the item properties of the items that make up the scale and thus can be different for low as compared to high scores.

Some of the practical advantages of using IRT models such as the GRM are that (1) the measurement precision of scale scores can be described conditional on the latent variable score and that (2) a person's latent variable estimate is based on weighted item responses and the relation of the response with the item properties. In CTT most often a person's score on a construct is simply the sum of item responses. Given that IRT model assumptions are met, the latent trait estimate is a better indicator of a person's true level on the trait continuum than CTT's summed scale score (Dumenci & Achenbach, 2008; Thomas, 2011b).

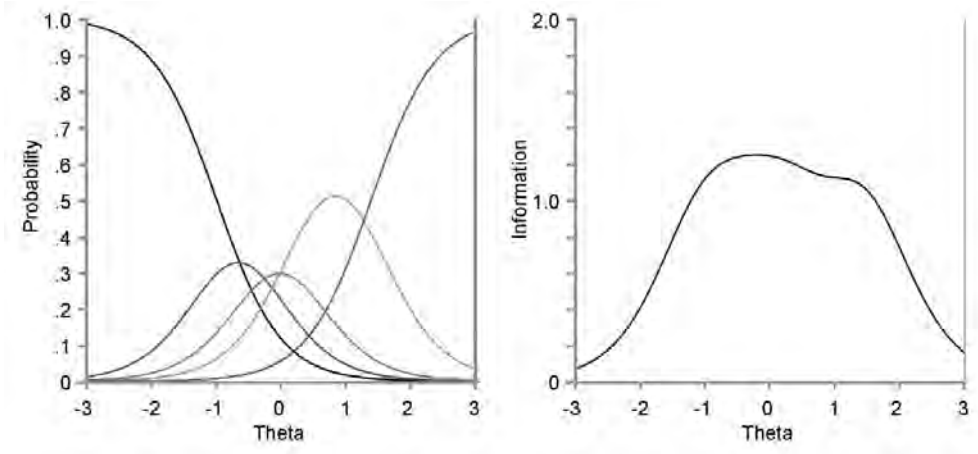

Figure 1.2: The CRFs and IIC for Item 7 'Introduce myself to new people' of the socially avoidant (FG) subscale of the IIP-64 in the sample of $N = 2263$ clinical outpatients from Chapter 3. The CRFs on the left side depicts the probability of a response in a particular response category conditional on the latent variable. The IIC on the right side shows that this item provides more information about persons that score above average on the socially avoidant scale as compared to those who have lower scores. Information is inversely related to the standard error.

1.5.2 The Bifactor Model

One of the assumptions underlying IRT models is that each item in a scale is influenced by a single unidimensional variable. However, due to their multidimensional nature psychological scales in clinical psychology are seldom strictly unidimensional. The question then is: Should we interpret groups of correlating items within a scale as separate scales or do all items have so much in common that we should interpret the scale as unidimensional? To answer this question, Reise et al. (2007; 2010) recommended researchers to complement the analyses of different correlated-trait factor models with a bifactor model. In a bifactor model each item loads on a general factor and is also allowed to load on each of the two or more orthogonal group factors. The general factor explains the item intercorrelations for all items, and, in addition, the group factors explain the item intercorrelations that attempt to capture the residual variation due to secondary dimensions. Figure 1.3 shows an example of a correlated-trait factor model on the left side and the corresponding bifactor model on the right side. In the bifactor model the item intercorrelations are interpreted as one common factor. With the bifactor model it is possible to (1) investigate the relation of items with the general factor and (2) to investigate how much variance of factors is unique once the common factor is already accounted for.

The bifactor model is strongly related to IRT because IRT researchers can use the bifactor model to check whether the data is unidimensional enough to meet the IRT assumptions. The bifactor model is also interesting with regard to two issues of this thesis. It provides a new point-of-view to investigate (1) the factor structure of clinical scales and (2) the added value of subscales above the total scale score.

1.6 Outline of this Thesis

This thesis consists of the following chapters. In Chapter 2, the problem under investigation is the dimensionality of the Dispositional Hope Scale (DHS; Snyder et al., 1991). Because researchers have made different recommendations with regard to the dimensionality of the DHS, it is unclear whether the use of subscale scores can be defended. The chapter describes the analyses of a one-factor model, a two-factor model, and a bifactor model for three samples: a student sample, a sample of psychiatric inpatients, and a sample of delinquents. In Chapter 3, I evaluate the psychometric quality of the eight subscales of the Inventory of

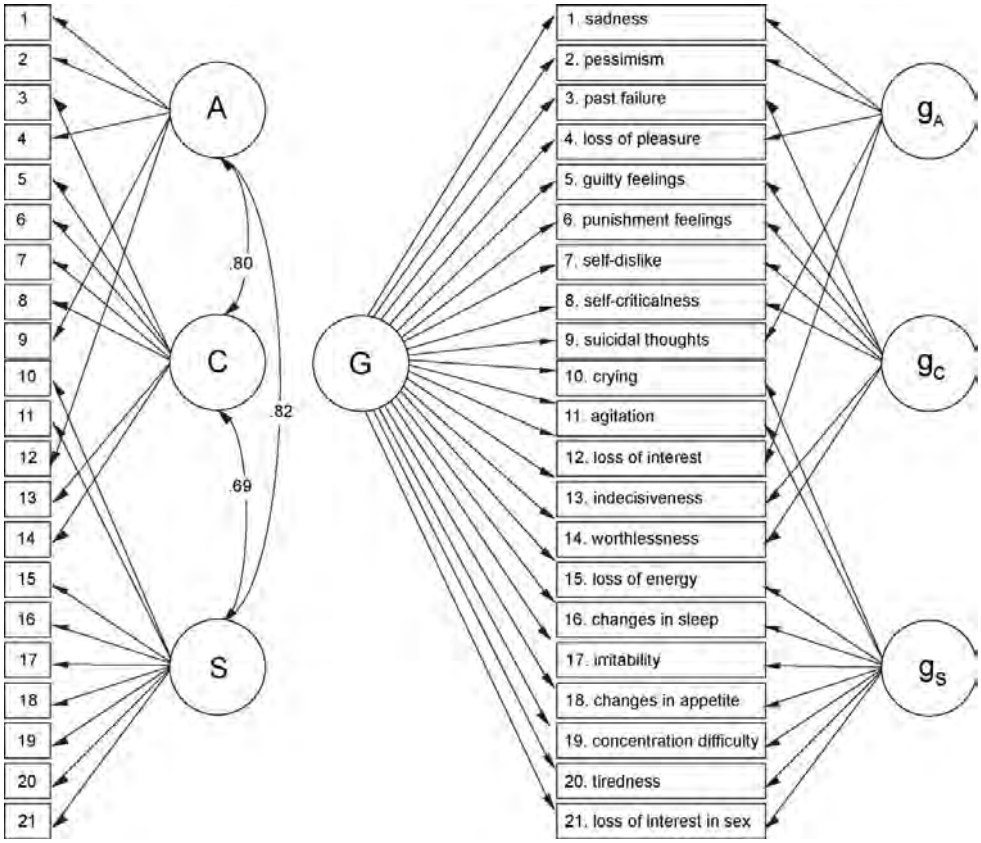

Figure 1.3: Correlated-traits factor model and bifactormodel for the BDI-II. The model on the left side is the correlated-traits three factor model for the BDI-II from the Beck et al. (2002) study, the model on the right side is the corresponding bifactor model with three group factors. C = Cognitive, S = Somatic, A = Affective, G = General factor, g_x are group-factors for the bifactor models.

Interpersonal Problems 64 (IIP-64; Horowitz et al., 2000). It is unclear how well the IIP-64 subscales tap the entire range of the underlying interpersonal problems dimension. I used results from different IRT analyses to investigate the reliability of subscale scores for different ranges of scale scores. Chapters 4 and 5 present two studies on the psychometric quality of the Beck Depression Inventory II (BDI-II; Beck et al., 1996). In Chapter 4, I first describe the discussion in the research literature about the dimensionality of the BDI-II. Second, I use bifactor analysis to answer the question whether BDI-II data are unidimensional enough to scale persons according to their depression scores and I compare results from a one-factor model and different two-factor, three-factor, and bifactor models in

a large sample of clinical outpatients. In Chapter 5, I analyze the measurement precision of the BDI-II scale using pre- and post treatment scores in a sample of clinical outpatients. Then, these results are used to discuss how GRM analyses of the BDI-II and an IRT-based change index can contribute to our understanding of the reliable measurement of individual change. Finally, in Chapter 6 of this thesis I tie the research findings from these four studies together to address the issues that were identified in this introduction and are related to the use of subscales, dimensionality of clinical scales, measurement precision, and the added value of IRT and bifactor analyses in our thinking about assessment in the field of clinical psychology.

All data that are used in this thesis are archival data. The chapters in this thesis are self-contained and can be read separately. Therefore, some overlap between the content of the chapters could not be avoided.

Chapter 2
On the Dimensionality of the Dispositional Hope Scale

Abstract

The Dispositional Hope Scale (DHS; C. R. Snyder et al., 1991) consists of two subsets of items measuring Agency and Pathways. The authors used bifactor analysis to evaluate the dimensionality structure of the scale. Data from 676 persons (295 psychiatric patients, 112 delinquents, and 269 students) were analyzed. The authors conclude that although the Pathway items seem to explain some additional variance when the Hope scale variance is partionalized out, the DHS allows unidimensional measurement.

This chapter has been published as:

Brouwer, D., Meijer, R. R., Weekers, A. M., & Baneke, J. J. (2008). On the dimensionality of the dispositional hope scale. *Psychological Assessment*, *20*, 310-315.

2.1 Introduction

Researchers and theorists within the positive psychology movement have devoted a great deal of energy to the study of human strengths (Seligman, 2005). The construct of hope (Snyder, 2000, 2004) has received increasing attention. It has been shown that hope is positively associated with positive affect, self-esteem (Snyder et al., 1991), and mental and physical health (Magaletta & Oliver, 1999) and negatively associated with depression and anxiety (Arnau et al., 2007; Chang, 2003; Snyder et al., 1991), feelings of burnout, and negative affect in general (Snyder et al., 1991).

An often-used instrument to measure hope is the Dispositional Hope Scale (DHS; Snyder et al., 1991). This 12-item scale consists of four items measuring Pathways, four items measuring Agency, and four filler items. Hope is formally defined as "a reciprocally derived sense of successful agency (goal-directed determination) and pathways (planning of ways to meet goals)" (Snyder et al., 1991, p. 571). Agency and pathways are not synonymous, and both components are necessary for hopeful thoughts (Snyder, 2000). In order to be hopeful about attaining their goals, people must be convinced that they are able to generate pathways to reach their goals (e.g., "I'll find a way to get this done") and have the perception that they can begin and continue movement along these imagined pathways (e.g., "I can do this"). These thoughts are called pathways and agency thoughts, respectively.

Although several studies have investigated the factorial structure of the DHS (discussed below), it is still unclear when we account for the general factor, Hope, how much unique variance is explained by the subdomains, Agency and Pathways. This is a crucial question because it may shed some light on whether it is useful to devote further research to discriminant and incremental validity of the Pathways and Agency component, as suggested in the literature. Another shortcoming of the existing literature is that the samples used in the various studies almost uniquely consisted of undergraduates. Because students do not manifest sufficient heterogeneity with respect to the hope trait, it is questionable whether the estimated correlations on the basis of which the factor structure is estimated are representative for other groups. The aim of this study was to investigate the unique contribution of the Agency and Pathways items when using more heterogeneous samples.

2.1.1 Studies on the Factor Structure of the DHS

Various studies have investigated the reliability and factor structure of the DHS. Snyder et al. (1991) found internal consistencies between .74 and .84 for the Hope scale and test–retest reliabilities ranging from .73 to .85 in different Caucasian samples, and the Agency and Pathways factors were positively correlated in each sample (*r*s ranged from .39 to .57). For college student samples, Snyder et al. (1991) used principal component factor analyses, each item loaded on its respective factor, but for the psychological treatment samples, this distinction was not evident. Factor loadings for two Pathways items were similar or higher on the Agency factor than on the Pathways factor. Babyak et al. (1993) conducted a confirmatory factor analysis on the DHS using four large samples of college students. The two-factor model fitted the data significantly better than did a one-factor model representing general Hope.

Recently, Roesch and Vaughn (2006) conducted a confirmatory factor analysis to investigate the factor structure of the DHS in a large, multiethnic sample. They found that a two-factor representation of the DHS fitted the data significantly better than did a one-factor model, although the interfactor correlation was large (r = .823). Furthermore, multigroup analyses revealed that factor scores were invariant across gender and ethnic groups. Although the two-factor model fit the data better than did a one-factor model, the more interesting question is, when does this multidimensionality of the two content facets interfere with the scaling of individuals on the common construct of hope? Any scale that is not simply the repeating of the same item over and over is going to have some multidimensionality.

Thus, although research showed that the two-factor model fits the data better than does a one-factor model, both Snyder et al. (1991) and Babyak et al. (1993) have suggested that a higher order factor model is the best predictor of outcome variables. Roesch and Vaughn (2006) also concluded that '*it is extremely difficult to predict outcome variables of interest from the Agency and Pathways factors because of the substantial overlap in variability [and] … on a theoretical and conceptual level these two hope constructs are described as relatively distinct. However, on a measurement level it is not clear that participants perceive measures of these constructs in this way'* (p. 82).

Interesting in this respect is a study by Arnau, Rosen, Finch, Rhudy, and Fortunato (2007), who tested the effects of the Agency and Pathways components of hope on depression and anxiety at three time points using a longitudinal study and cross-lagged panel models. Results showed significant negative effects for the Agency component of hope on later depression but no unique effect of the Pathways component of hope on depression. Likewise, Agency showed a negative effect on later anxiety, but Pathways had no significant influence on anxiety. However, as Arnau et al. (2007) discussed, these outcomes may result from the shared variance of both Agency and Pathways rather than from each making independent contributions to the relationships. If Agency and Pathways are both accounting for the same variance in depression and anxiety, then one latent variable would receive credit for the effect with a statistically significant cross-lag parameter estimate in the structural model and the other would receive a statistically insignificant parameter estimate. However, Chang (2003) found a strong difference on pathways thinking in middle-aged men compared to women and therefore noted that, although most studies examining Snyder's hope theory have been based on using the total Hope score, it is important to distinguish between Agency and Pathways items.

We conclude from the studies cited above that it is still unclear how much unique variance is explained by the subdomains Pathways and Agency above the general Hope component. The study by Arnau et al. (2007) suggested that when both components are treated as separate dimensions, but in reality share much common variance, results from validity studies cannot be trusted. Therefore, in the present study, we further investigated the dimensionality structure of the DHS. To investigate the psychometric structure of the DHS, we used confirmatory factor analysis. Aside from unidimensional and multidimensional models, we used the bifactor model (Chen, West, & Sousa, 2006; Holzinger & Swineford, 1937; Reise, Morizot, & Hays, 2007), which has not been applied before to evaluate the DHS. In a bifactor model, there is a general factor that explains the item intercorrelations, but in addition, there are also so-called group factors that explain the item intercorrelations that attempt to capture the residual variation due to secondary dimensions. Thus, in the case of the DHS, the question is, how much variation is unexplained when the factor Hope is already taken into account? The bifactor model can be particularly useful in testing whether a subset of the domain-specific factors predicts external variables, over and above the general factor.

2.2 Method

2.2.1 Measures and Participants

To measure hope, we used a Dutch version of the DHS (translated by Joost J. Baneke; Snyder et al., 1991) with the original 8-point Likert-type scale. The original version of the DHS is depicted in Table 2.1.

Table 2.1

Hope Scale Items and Lower-Order Level Domains.

Item number	Domain	Item Content
1	P	I can think of many ways to get out of a jam
2	A	I energetically pursue my goals
		I feel tired most of the time
3	P	There are lots of ways around any problem
		I am easily downed in an argument
4	P	I can think of many ways to get the things in life that are most important to me
		I worry about my health
5	P	Even when others get discouraged, I know I can find a way to solve the problem
6	A	My past experiences have prepared me well for my future
7	A	I've been pretty successful in life
		I usually find myself worrying about something
8	A	I meet the goals that I set for myself

Note: P = Pathways, A = Agency, filler items are printed in italics

The sample included 676 persons, of whom 295 were psychiatric patients (107 men, 188 women), 112 were delinquents (102 men, 10 women), and 269 were students (79 men, 190 women). Mean ages were 30.7 years ($SD = 11.3$) for men and 25.8 years ($SD = 8.4$) for women. Data from delinquents and psychiatric patients were obtained as part of a psychological assessment program; student data were collected for research purposes. We conducted two analyses. In the first analysis, we used all 676 persons, so that we analyzed a heterogeneous sample. A possible drawback of this approach is that, because of the mixed population, this may lead to inflated correlations. Therefore, we also analyzed the data for the

students and psychiatric patients separately. It was not possible to run the analysis for the delinquents only because of the small sample size.

In contrast to earlier studies on the DHS, we analyzed different samples of persons. Almost all earlier studies used a convenience sample of students. However, it is well known that to obtain replicable factors, researchers should assemble samples with sufficient person representation at all levels of the trait dimensions (see Reise, Waller, & Comrey, 2000). One consequence is that using a sample of students may be suitable when students manifest sufficient heterogeneity with respect to the trait standing. On some constructs, such as extraversion or agreeableness, this seems reasonable. For a construct like hope, however, students may not be an appropriate respondent pool to accurately map the factor space of a clinical assessment scale because most students will have relatively high scores on the hope construct. By including psychiatric patients and delinquents in the analysis, we obtain more heterogeneity with respect to the hope trait. For the samples used in this study, we found differences in mean scores on the DHS for psychiatric patients (M = 33.89, SD = 11.39), delinquents (M = 42.69, SD = 11.14), and students (M = 47.32, SD = 6.49), where higher scores point at a more hopeful attitude.

2.2.2 Analyses

Thus far, researchers have used a one-factor model, a two-factor model, and a second-order model (i.e., a model with items loading on first-order factors, Pathways and Agency, and first-order factors loading on the second-order factor, Hope) to analyze the DHS. In this study, we analyzed the DHS using the one-factor, two-factor, and bifactor models. In the one-factor model, all items load on one general Hope factor. In the two-factor model, each item only loads on one out of two factors, Agency or Pathways, and the factors may be correlated. In the bifactor model (see Figure 2.1), each item has a loading on the general Hope factor and on one of the group factors, Agency and Pathways. It is important to understand that although the second-order model and the bifactor model are not equivalent, they have similar interpretations. Chen et al. (2006) discussed differences and similarities between the two models. We discuss three important similarities and differences. First, the second-order factor (Hope) in the second-order model corresponds to the general factor in the bifactor model. Second, the disturbances of the first-order factors in the second-order model resemble the domain-specific factors in the bifactor model. The advantage of the bifactor model, however,

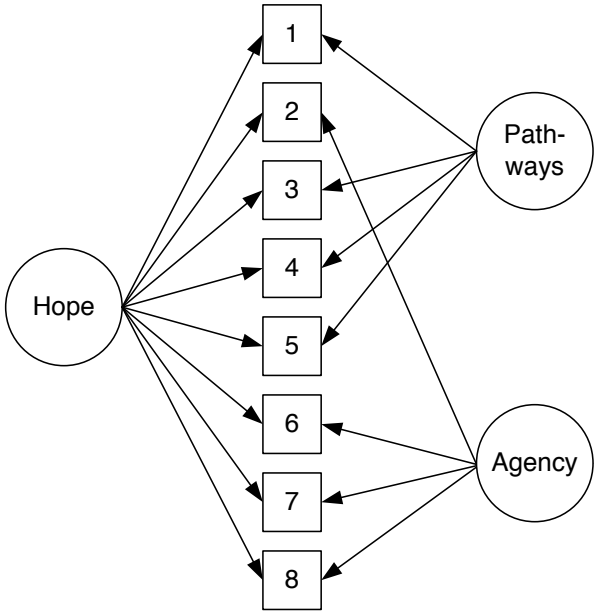

Figure 2.1: Bifactor model for the DHS scale

is that it can be used to study the role of domain-specific factors that are independent of the general factor.

To illustrate this, consider the DHS. The factor of Hope is of focal interest, as are two domain-specific factors of hope, Agency and Pathways. Suppose now that Agency reflects only hope, whereas Pathways still exists as a specific domain even after partialling out the general Hope factor. In this example, Agency will not exist as a domain-specific (i.e., lower order) factor in the bifactor model, but it will exist in the second-order model. Pathways will exist as a domain-specific factor in the bifactor model and as a lower order factor in the second-order model. The lack of significance in the variance of the disturbance will typically not cause any problem in a second-order factor model, and therefore, the possibility that one domain-specific factor may not exist can be easily overlooked.

A third important advantage of the bifactor model is that we can directly examine the strength of the relationship between the domain-specific factors and their associated items. The relationship is reflected in the factor loadings, whereas the relationship cannot be directly tested in the second-order factor model, as the domain specific factors are represented by

disturbances of the first-order factors. Thus, although Babyak et al. (1993) presented hope as an overarching construct, we ask ourselves when we account for the general factor (Hope), how much unique variance is explained by the subdomains? In other words, how much variance do Agency and Pathways share, and how much variance is unique? Also, the higher order factor "emerges" from the correlation among the subfactors; that is, it explains subscale correlations, while in the bifactor model, the general factor is a latent variable that explains item correlations. Only in the bifactor model can item variance be partitioned into that due to the general and that due to the group factors.

The confirmatory one-factor, two-factor, and bifactor models were estimated using MPLUS 4.1 (Muthén & Muthén, 1998 – 2006). The maximum likelihood estimation option was used for all calibrations, and consequently, for model evaluation the program provides a likelihood ratio χ^2, the number of free parameters, and three information criteria, namely, the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the sample-size adjusted BIC (ABIC). We also used weighted least squares (WLS) estimation. Using WLS estimation, the one-factor, two-factor, and bifactor models and their fit statistics (comparative fit index [CFI], Tucker-Lewis index [TLI], root mean square error of approximation [RMSEA], and standardized root mean squared residual [SRMR]) are provided. However, due to computational difficulties, WLS estimation was not possible for the psychiatric patient and student samples. Therefore, we only report CFI, TLI, RMSEA, and SRMR fit indices for the complete sample.

2.3 Results

Tables 2.2, 2.3, and 2.4 display the mean item scores, classical item–test correlations, and factor loadings under the unidimensional model, the multidimensional model, and the bifactor model for the total group (see Table 2.2), the psychiatric patients (see Table 2.3), and the students (see Table 2.4), respectively.

In terms of model evaluation for the complete sample, we found a good fit under WLS estimation for all three models; TLI and CFI are higher than .95, and RMSEA is close to .06 for the multidimensional and bifactor models, and SRMR is lower than .08 for all three models. With respect to the information criteria, the results were not straightforward. The AIC is reduced when comparing the unidimensional model, the multidimensional model, and

Table 2.2

Results for the Confirmatory and Bifactor Analysis for the Hope Scale (Whole Sample).

			One factor analysis	Two factor analysis (r = .913)		Bifactor analysis		
Item	Mean value	Item-test	Hope	Agency	Path ways	Hope (G)	Agency	Path ways
1	5.18	.71	.794		.812	.738		.343
2	5.44	.62	.673	.700		.690	.393	
3	5.93	.58	.647		.670	.576		.413
4	4.94	.72	.814		.824	.770		.264
5	4.92	.61	.700		.713	.652		.291
6	5.01	.63	.692	.705		.712	-.040	
7	4.40	.60	.656	.675		.688	-.076	
8	4.87	.69	.762	.795		.787	.155	
LL			-9475.887	-9459.023		-9447.840		
FP			64	65		72		
AIC			19079.774	19048.047		19039.680		
BIC			19368.810	19341.599		19364.846		
ABIC			19165.604	19135.218		19136.239		
CFI			.987	.994		.996		
TLI			.982	.991		.991		
RMSEA			.106	.073		.075		
SRMR			.035	.025		.021		

Note. Factor loadings estimated under maximum likelihood estimation. G = general factor; LL = loglikelihood; FP = number of free parameters; AIC = Akaike information criterion; BIC = Bayesian information criterion; ABIC = sample-size adjusted BIC; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean squared residual.

the bifactor model. However, the BIC is smaller for the multidimensional model than for the bifactor model and the unidimensional model, whereas the ABIC is similar for the multidimensional model and the bifactor model and higher for the unidimensional model, indicating that a multidimensional model and a bifactor model give a better representation than does a unidimensional model. These findings also apply for the psychiatric patients and students separately (see Tables 2.3 and 2.4).

Table 2.3

Results for the Confirmatory and Bifactor Analysis for the Hope Scale (Psychiatric Patients).

			One factor analysis	Two factor analysis (r = .866)		Bifactor analysis		
Item	Mean value	Item-test	Hope	Agency	Path ways	Hope (G)	Agency	Path ways
1	4.31	.65	.763		.802	.666		.536
2	4.78	.59	.661	.705		.673	.318	
3	5.33	.57	.662		.692	.591		.384
4	3.88	.65	.770		.771	.735		.216
5	4.33	.58	.677		.699	.602		.352
6	3.97	.50	.592	.599		.624	-.133	
7	3.38	.50	.602	.635		.638	.025	
8	3.93	.65	.750	.797		.790	.204	
LL			-4321.028	-4309.577		-4301.483		
FP			64	65		72		
AIC			8770.055	8749.155		8746.966		
BIC			9006.021	8988.808		9012.428		
ABIC			8803.058	8782.674		8784.095		

Note. G = general factor; LL = loglikelihood; FP = number of free parameters; AIC = Akaike information criterion; BIC = Bayesian information criterion; ABIC = sample-size adjusted BIC.

Inspection of the factor loadings in Table 2.2 for the multidimensional model reveals that the items appear to be fairly good measures of their respective dimensions, but dimensions are highly correlated (r = .913). Comparison of the multidimensional model with the unidimensional model shows that the factor loadings are also high for the unidimensional scale and that the factor loadings show only minor increases when dividing the scale into two dimensions. When inspecting the bifactor model, it is clear that the Hope items are discriminating measures of the general factor and that the factor loadings are significantly higher for the general factor than for either of the two group factors, which is in concordance with the minor differences between the unidimensional and multidimensional models. Table 2.2 also shows that in the bifactor model, the loadings on the general factor for the Pathway items (Items 1, 3, 4, and 5 in Tables 2.2, 2.3, and 2.4) tend to go down only slightly relative to the loadings in the one-factor solution. However, for the Agency items (Items 2, 6, 7, and 8), the loadings on the general factor are similar or even a bit higher than in the one-factor

Table 2.4

Results for the Confirmatory and Bifactor Analysis for the Hope Scale (Students).

			One factor analysis	Two factor analysis (r = .761)		Bifactor analysis		
Item	Mean value	Item-test	Hope	Agency	Path ways	Hope (G)	Agency	Path ways
1	6.12	.55	.659		.687	.521		.446
2	5.92	.50	.577	.690		.657	.129	
3	6.55	.30	.422		.476	.203		.647
4	5.87	.61	.742		.780	.596		.481
5	5.43	.47	.585		.593	.500		.267
6	5.95	.46	.502	.507		.601	-.453	
7	5.68	.47	.482	.514		.498	.078	
8	5.80	.47	.574	.663		.719	.446	
LL			-3233.643	-3222.140		-3204.349		
FP			57	58		65		
AIC			6581.287	6560.281		6538.698		
BIC			6786.186	6768.774		6772.354		
ABIC			6605.459	6584.877		6566.262		

Note. G = general factor; LL = loglikelihood; FP = number of free parameters; AIC = Akaike information criterion; BIC = Bayesian information criterion; ABIC = sample-size adjusted BIC.

solution. Furthermore, note that the loadings on the group factors (Agency and Pathways) are approximately zero for the three Agency Items 6, 7, and 8 and around .30 for the Pathway items. Thus, the factor loadings are low to very low for most items when partialling out the common variance.

In Table 2.3, the results are depicted for the psychiatric patients. Results and trends for the psychiatric patients are similar to those for the total group, although factor loadings are somewhat lower. In the bifactor model, the loadings of most Agency and Pathway items are low when the common variance is partialized out. Only Item 1 (Pathways) explains some additional variance as compared to the general factor.

For the students (see Table 2.4), the factor loadings are (sometimes considerably) lower than for the psychiatric patients for the unidimensional and multidimensional models. For the bifactor model, we found higher loadings on the Pathway items for students than for the

psychiatric patients. Furthermore, Item 3 had a higher loading on the Pathway factor than on the General factor. However, because the scale is not very suited to discriminate students from each other (as suggested by the relatively low factor loadings), these results are less informative than are the results for the psychiatric patients.

In general, we conclude that although there seems to be some additional variance above the Hope factor in the Pathway items (there is none for the Agency items), this additional variance seems to be very small and does not seem to justify a separate treatment of the Pathways and Agency items.

2.4 Discussion

We investigated the psychometric structure of the DHS and found that the best choice is to consider the scale as a unidimensional scale. These results were confirmed by a preliminary analysis we conducted for the complete sample. Using Mokken scale analysis (e.g., Meijer & Baneke, 2004; Sijtsma & Molenaar, 2002), we found that the DHS formed a strong scale. Also, exploratory factor analysis resulted in a first eigenvalue of 4.47 and a second eigenvalue of 0.74. Thus, the ratio of first to second eigenvalues equaled 6.05, which pointed to a strong common factor.

In earlier studies about the structure of the DHS, Roesch and Vaughn (2006) suggested that on a measurement level, it is not clear that participants perceive Agency and Pathways as distinct constructs. Arnau et al. (2007) suggested that Agency and Pathways do not necessarily make unique, independent contributions to the Hope construct. They did not find that Pathways uniquely predicted depression and anxiety. On the basis of our results, we do not think that it will be very fruitful, as suggested by Snyder et al. (1991), to unravel differential correlates of Agency and Pathways yielding information pertaining to their separate construct validity and utility. Our results showed that the items measure the same construct and that there is very little unique variance that is explained by the Pathways or Agency items above the general factor Hope.

Chapter 3

The Psychometric Quality of the Individual Scales of the Inventory of Interpersonal Problems 64: An Item Response Model Approach

Abstract

This study evaluated the psychometric quality of the eight subscales of the IIP-64. Both nonparametric and parametric item response theory models were used to identify the relative effectiveness of items in discriminating between levels of interpersonal distress within the specific subscales and to obtain information about the standard error of measurement for different subscale scores in a large sample (N = 2236) of clinical outpatients. Five of the IIP-64 subscales form scales of medium quality; for three subscales the items are unscalable. Measurement precision differed across the latent trait ranges for all scales. We conclude that when using IIP-64 subscales in, for example, outcome measurement scales should be used with care because items do not tap the entire range of severity and three subscales do not allow precise measurement.

This chapter has been submitted as:

Brouwer, D., Meijer, R. R., & Zevalkink, J. The Psychometric Quality of the Individual Scales of the Inventory of Interpersonal Problems 64: An Item Response Model Approach.

3.1 Introduction

The Inventory of Interpersonal Problems 64 (IIP-64, Alden, Wiggins, & Pincus, 1990; Horowitz et al., 2000) is one of the most frequently used and well-established psychological inventories to assess interpersonal problems in clinical treatment centers and in research applications. This self-report instrument is intended to determine the amount and type of problems persons experience in relating to significant persons in their life. The IIP-64 consists of eight subscales representing eight domains of interpersonal behavior, each consisting of eight items: Domineering (PA), Vindictive (BC), Cold (DE), Socially Avoidant (FG), Nonassertive (HI), Exploitable (JK), Overly Nurturant (LM), and Intrusive (NO)[1]. The psychometric properties of the IIP-64 have been widely investigated. For example, the factor structure was investigated by Acton and Revelle (2002), Grosse-Holtforth, Lutz, and Grawe (2006), Pincus, Gurtman, and Ruiz (1998), Tracey, Rounds, and Gurtman (1996), and Vanheule, Desmet, and Rosseel (2006), its sensitivity for detecting change by Huber, Henrich, and Klug (2007), its usefulness in relation to other measures of interpersonal behavior by Alden et al. (1990), Horowitz et al. (2000), Leising, Rehbein, and Sporberg (2007), and Vittengl, Clark, and Jarrett (2003), and its relation to psychotherapy outcome by Horowitz, Rosenberg, and Bartholomew (1993), Puschner, Kraft, and Bauer (2004), Ruiz, Pincus, Borkovec, Echemendia, Castonguay, and Ragusea (2004) and Vittengl et al. (2003).

Researchers use the IIP-64 scale as a whole representing interpersonal distress, or use the subscales to differentiate between the specific domains of interpersonal functioning. In clinical practice, IIP-64 subscale scores are used to compare an individual person or group of persons with a normative sample or to compare a person's distress in each interpersonal domain relative to the person's overall level of interpersonal distress (so-called ipsatized scores) which allows the clinician to identify domains that the individual experiences as particularly problematic, regardless of the person's overall reported level of interpersonal problems. Ipsatized subscale scores can be ordered around a circle located on a two-dimensional graph. The two axes of the graphs correspond to a dimension of affiliation and a

[1] The abbreviations of the subscales originate from sixteen (A-P) positions counterclockwise on a circular (circumplex) structure.

dimension of dominance. The theoretical and methodological underpinnings of the circumplex factor structure of the IIP-64 have been the focus of many research studies (e.g., Horowitz et al., 2000; Acton & Revelle, 2002; Pincus et al., 1998; Tracey et al., 1996; Vanheule et al., 2006).

Significantly less research has been conducted to determine the psychometric quality of the individual subscales, such as the abilities of the items and subscales to differentiate between individuals with different severity of specific interpersonal problems, and the measurement precision for different individual subscale scores. Detailed information about the psychometric quality of the subscales is an essential prerequisite to a sensible use of these subscales in clinical practice (e.g., comparison with norm sample, or use of ipsatized subscale scores) and in research (e.g., to study the circumplex factor structure, or to measure therapeutic change and outcome).

Several aspects of the psychometric quality of the IIP-64 subscales have been investigated. Many studies reported reliability estimates of the subscales. Coefficient alpha for the eight IIP-64 subscales ranged from .65 through .88, with NO having the lowest estimated reliability (between .65 and .73) and FG and HI subscales having the highest estimated reliability (between .82 to .88; see Grosse-Holtforth et al., 2006; Horowitz et al., 2000; Leising et al., 2007; Vanheule et al., 2006; Vittengl et al., 2003). Furthermore, Vanheule et al. (2006) conducted one of the few studies that investigated the unidimensionality of the subscales of the IIP-64. A confirmatory factor analysis of covariance matrices, a priori specifying an eight factor model, demonstrated a bad global model fit for the eight subscales. They concluded that more research is needed to further investigate the quality of the IIP-64 subscales.

Recently, Doucette and Wolf (2009) questioned the psychometric quality of many instruments used in psychotherapeutic research and practice and advocated the use of item response theory (IRT, Embretson & Reise, 2000) to obtain more detailed information about these instruments. Also, Thomas (2012) concluded in a review of the usefulness of IRT in clinical assessment that "*IRT has the potential to drastically alter test selection, model development, and scoring*" (p. 13). These authors describe the advantages of using IRT in obtaining other sources of information about the quality of scales as compared to only using classical test theory. For analyzing the quality of the IIP-64 subscales, IRT has two important advantages to classical scale analyses. By means of IRT modeling it is possible (1) to identify

the relative effectiveness of items in discriminating between levels of interpersonal distress within the specific subscales, and (2) to obtain information about the standard error of measurement of an IIP-64 subscale for different subscale scores.

We would like to stress that it is still unclear how well IIP-64 subscales tap the entire range of the underlying interpersonal problem continuum. Reliable scales should include items that tap the entire range of severity in order to differentiate between individuals and to be sensitive to change over time. The aim of the present study is to contribute to previous research, by using IRT techniques with data from a large sample of clinical outpatients, to obtain new information about the subscale quality of the IIP-64.

3.2 Method

2.3.1 Measures

The IIP was originally developed to map individual differences with respect to several domains of interpersonal relations that were problematic for patients undergoing some form of psychotherapy (Horowitz, Rosenberg, Baer, Ureño, & Villaseñor, 1988; Horowitz & Vitkus, 1986). The original IIP consisted of 128 questions related to reported problems in the interpersonal domain in psychotherapeutic sessions, and was reduced to the widely-used IIP-64, also named the IIP-Circumplex (IIP-C, Alden et al., 1990), and later shortened into three different IIP-32 versions (Barkham, Hardy, & Startup, 1996; Horowitz et al., 2000; Soldz, Budman, Demby, & Merry, 1995). For the IIP-64, Alden et al. (1990) selected subscale items so as to maximize the subscales' fit to a circumplex structure. In the circumplex structure subscales are arranged in a circular array in a two dimension space, such that subscales that are close together are more positively related than subscales that are further apart on the circle (for a more extensive description of psychometric criteria of the circumplex structure, see for example Acton and Revelle, 2004; Browne, 1992). The IIP-64 questions are rated on a 5-point scale (0 = 'not at all', 1 = 'a little bit', 2 = 'moderately', 3 = 'quite a bit', 4 = 'extremely'). In each item a respondent has to consider how distressful a particular interpersonal problem has been with respect to any significant person in their life (e.g., "I try to change other people too much"). For each domain a scale score is obtained by calculating the sum of the eight item responses that represent the domain. A high score on each scale is indicative for problematic interpersonal behavior in a particular domain.

2.3.2 Participants

The sample included N = 2263 outpatients, 39.9% males and 60.1% females. Mean age was 34.1 years (SD = 10.2) for females and 36.8 years (SD=10.3) for males. Data were obtained as part of the psychological screening assessment procedure for persons applying for treatment in the period between the years 2004 and 2009 at a community mental health clinic specialized in ambulatory psychoanalytic treatment. Participants were selected that gave informed consent and completed the IIP-64. Psychiatric diagnoses were assessed by mental-health specialists in a consensus meeting and classified according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-R; American Psychiatric Association, 2000). For 93.8% of the 2263 participants DSM-IV-R Axis I classifications were available. Most frequent classifications were mood disorders (48.8%), identity problems (31.0%), partner relational problems (25.6%), anxiety disorders (17.8%), phase of life problems (11.5%), adjustment disorders (10.4%), and substance related disorders (10.2%). Axis II classifications were not assessed systematically. 90.4% of the outpatients reported Dutch as their dominant culture when asked in which culture they were raised and 62.3% of the outpatients had a Bachelor or Master degree.

2.3.3 Analyses

Item Response Theory. Several authors discussed the advantages of applying IRT models to construct and to investigate personality and mood disorder scales (e.g., Reise & Henson, 2003; Meijer, Egberink, Emons, & Sijtsma, 2008). We will summarize the most important assumptions and advantages of IRT that are relevant to the current study.

IRT models are based on the idea that psychological constructs are latent, that is, not directly observable, and that knowledge about these constructs can only be obtained through the manifest responses of persons to a set of items (e.g., Embretson & Reise, 2000; Sijtsma & Molenaar, 2002). IRT explains the structure in the manifest responses by assuming the existence of a latent trait, denoted by the Greek letter θ. By means of IRT models it is possible to locate a person's θ and the characteristics of the items that make up the measurement instrument, on the same metric (i.e., latent trait continuum). For dichotomous items, unidimensional IRT is based on the assumption that a person's performance on a test item can be predicted by the interplay between θ and item characteristics such as item

discrimination and item difficulty (e.g., Embrteson & Reise, 2000). The relationship between item performance and the trait level θ can be described by a monotonically increasing function, which is called the item response function (IRF). Let $P_i(\theta)$ be the probability of a positive response (i.e., a correct answer or the agreement with a specific statement) on item i for a given level of θ. Then the core assumption states that when the trait level θ increases, the probability of a positive item response $P_i(\theta)$ also increases. For polytomously scored items, this assumption is made at the level of item steps, which are the transitions from one answering category to the next. For example, respondents choosing category 2 on a four-point scale have a score of 1 on the first two item steps (from 0 to 1 and from 1 to 2) and a score of 0 on the second two item steps (from 2 to 3 and from 3 to 4).

In IRT, nonparametric and parametric approaches can be distinguished. Nonparametric IRT models are based on less restrictive assumptions about the data and are, therefore, ideal instruments to explore the psychometric structure of tests. Parametric approaches are based on more restrictive assumptions, but provide analytical tools, such as information functions, that cannot be obtained using nonparametric approaches. In this study we used Mokken's nonparametric monotone homogeneity model (MMH; Sijtsma & Molenaar, 2002; Meijer & Baneke, 2004) to explore the psychometric structure of the IIP-64 scales and the parametric graded response model (GRM; Samejima, 1969, 1997) to obtain more detailed information about the measurement precision of the IIP-64 subscales across different latent trait values. Furthermore, we used both approaches to obtain a detailed picture about the psychometric quality of the subscales.

Mokken scaling. We used the computer program Mokken Scale Analysis for Polytomous Items version 5.0 (MSP5.0; Molenaar & Sijtsma, 2000) to conduct a Mokken scale analysis for each scale of the IIP-64. The model assumes that all items in a test measure the same latent trait (unidimensionality assumption), that a person's response to one items is not influenced by the response to another item (local independence), and that the item response function is nondecreasing (monotonicity assumption). A more detailed description of these assumptions can be found in Sijtsma and Molenaar (2002) or Meijer and Baneke (2004).

We calculated the coefficient H_i for items and the coefficient H for a set of items to check the scalability of the items, that is, the degree to which a set of items are related to each other and form a scale. Under the MMH, higher positive H values reflect higher discrimination power

of the items, and as a result, more confidence in the ordering of respondents by means of their total scores. Items with high H_i values discriminate well in the group in which they are used. H_i values determine how well an item fits the scale. For practical test construction purposes, the following rules of thumb have been suggested. Weak scalability is obtained if $.3 \leq H \leq .4$, medium scalability if $.4 \leq H \leq .5$, and strong scalability if $.5 \leq H < 1$ (Sijtsma & Molenaar, 2002, pp 60-61). If $H < .30$ it would be misleading to conclude that measurements on such a scale discriminate between persons. Furthermore, we checked the monotonicity assumption by inspecting the item step response functions. That is, we checked the graphs of the proportion of positive response per item step conditional on the rest scores. The steep- or flatness of the graphs indicates the ability of the item to differentiate between persons with low, average, or high rest scores. For the IIP-64 subscales the rest score is the total score on a subscale minus the item score, and thus gives an indication of the severity of interpersonal distress for that specific domain.

The graded response model. The GRM is suitable for analyzing ordered response categories, such as likert-type rating scales. Several researchers used this model to analyze personality data and there is a close relationship between the GRM and Mokken's MMH model (Sijtsma & Molenaar, 2002, p. 129). The items in the GRM are defined by a discrimination parameter (a; usually with numerical values between .5 and 2.5) and two or more location parameters (b; numerical values between -2.5 and 2.5); the number of location parameters per item is equal to the number of response categories minus 1, in our analysis 5-1 = 4. Like the H-coefficient, the magnitude of the discrimination parameter reflects the degree to which the item is related to the underlying latent trait. This means that for high a-values the response categories accurately differentiate among trait levels. The location parameter b_m can be interpreted as the point at the latent trait continuum where there is a 50% chance of scoring in category m or higher. Thus, respondents with a θ-value higher that b_3 have more that 50% chance of responding in category 3 ('quite a bit') or higher (4 = 'extremely').

An important difference between Mokken scaling and the GRM is that in the former, persons are assumed to have equal standard errors regardless of their position on the construct. In Mokken scaling, like in classical test theory, there is one reliability estimate. In parametric IRT, the concept of reliability is replaced by the concepts of item and scale information. The standard error of a trait estimate is inversely related to the square root of the test information

function. Thus, persons may have different standard errors depending on how discriminating a set of items is in different ranges of the latent trait. In general, items with larger discrimination parameters (i.e., the *a* parameters) provide relatively more information. The location parameters (i.e., the *b* parameters) determine where the information is located. Item information is additive across the items administered and test information is maximized around the location parameters. The item and scale information curves graphically show the information conditional on the latent trait. Because information is inversely related to the standard error of measurement this feature of IRT allows us to determine how precise a measure is for individuals in high, medium, and low trait ranges. We estimated the items discrimination and location parameters for the GRM using Multilog 7.0 (Thissen, Chen & Bock, 2003).

3.3 Results

3.3.1 Descriptive statistics and nonparametric scaling

Table 3.1 contains the mean values, standard deviations, item-total correlations, and *H*-values for items and subscales. It also contains the item parameters estimated under the graded response model, to be discussed below. A first observation is that the mean values are skewed to the right (note that, theoretically, subscale scores range from 0 to 32). Mean values are lowest for the subscales DE (M = 8.95), PA (M = 8.12), and BC (M = 7.49), and highest for the subscales HI (M = 14.72), JK (M = 14.08), and LM (M = 14.23). These mean total scores for our sample of clinical outpatients are similar as those found in other clinical samples (e.g., Horowitz et al., 2000; Puschner et al., 2004; Vittengl et al., 2003).

Cronbach's α ranges from α = .70 through α = .85 and item-total correlations range from r_{it} = .19 to .73. Overall *H*-values range from .24 through .44 indicating weak to medium scales. Five of the IIP-64 subscales form medium scales with $H \geq .40$, but the items of the three subscales PA (H = .30), BC (H = .28), and NO (H = .24) are unscalable, several items of these scales have H_i values smaller than H_i = .30. In general, H_i values range from H_i = .12 through H_i = .53 across all IIP-64 items. Sixteen items with $H_i < .30$ are unscalable; most of these items are from the PA, BC, and NO scales. These items do not discriminate well between persons with different latent trait values.

Table 3.1

Descriptive Statistics and IRT Parameters for the IIP-64 Items.

		Descriptive and Mokken statistics				GRM parameters				
Nr	Item content	*M*	*SD*	r_{it}	H_i	*a*	b_1	b_2	b_3	b_4
	Domineering (PA)									
45	I am too aggressive toward other people	.59	.97	.47	.32	1.38	0.64	1.51	2.33	3.93
57	I manipulate other people too much to get what I want	.70	1.02	.48	.32	1.52	0.35	1.27	2.09	3.40
17	Understand another person's point of view	.77	1.02	.43	.28	1.15	0.23	1.25	2.50	4.53
59	I argue with other people too much	.86	1.03	.55	.35	1.52	-0.01	0.97	2.04	3.61
50	I try to control other people too much	1.19	1.22	.53	.34	1.80	-0.38	0.51	1.23	2.43
52	I try to change other people too much	1.19	1.17	.53	.34	1.75	-0.46	0.51	1.33	2.67
44	I am too independent	1.29	1.27	.23	.16	0.52	-0.90	0.72	2.67	5.71
31	Take instructions from people who have authority over me	1.53	1.33	.39	.26	0.89	-1.07	0.16	1.26	2.97
Total		8.12	5.45	*H* = .30						α = .75
	Vindictive (BC)									
22	Be supportive of another person's goals in life	.49	.88	.39	.27	0.98	1.07	2.10	3.48	5.31
40	I fight with other people too much	.63	1.03	.35	.24	1.03	0.77	1.71	2.59	4.57
64	I want to get revenge against people too much	.64	1.04	.50	.32	1.64	0.55	1.,32	1.97	2.97
32	Feel good about another person's happiness	.79	1.06	.46	.30	1.15	0.22	1.26	2.32	4.17
24	Really care about other people's problems	.90	1.14	.43	.28	0.95	0.08	1.25	2.35	4.15
29	Put somebody else's needs before my own	1.19	1.27	.29	.20	0,60	-0,57	1.03	2.53	4.91
56	I am too suspicious of other people	1.24	1.24	.55	.35	2.10	-0.39	0.40	1.06	2.30
1	Trust other people	1.61	1.31	.42	.29	1.50	-0.92	0.06	0.74	2.14
	Total	7.49	5.30	*H* = .28						α = .73

(continued)

		Descriptive and Mokken statistics				GRM parameters				
Nr	Item content	*M*	*SD*	r_{it}	H_i	*a*	b_1	b_2	b_3	b_4
		Cold (DE)								
27	Give a gift to another person	.50	.92	.41	.34	1.19	0.97	1.83	2.74	
16	Get along with people	.91	1.10	.58	.42	1.75	-0.02	0.79	1.73	
23	Feel close to other people	.98	1.23	.65	.47	2.43	0.06	0.66	1.16	
20	Experience a feeling of love for another person	1.02	1.27	.60	.44	2.02	0.05	0.61	1.19	
36	Forgive another person after I've been angry	1.21	1.26	.67	.26	2.35	-0.29	0.38	1.05	
15	Show affection to people	1.21	1.25	.34	.48	0.75	-0.62	0.79	2.05	
11	Make a long-term commitment to another person	1.40	1.41	.58	.42	1.68	-0.38	0.26	0.83	
60	I keep other people at distance too much	1.72	1.33	.57	.43	1.61	-1.01	-0.13	0.59	
Total		8.95	6.60	*H* = .41						α = 83.
		Socially Avoidant (FG)								
7	Introduce myself to new people	.90	1.11	.53	.40	1.56	0.02	0.89	1.76	3.07
55	I am too afraid of other people	1.17	1.26	.53	.40	1.57	-0.26	0.53	1.24	2.57
33	Ask other people to get together socially with me	1.39	1.30	.60	.44	1.78	-0.54	0.25	0.94	2.12
14	Socialize with other people	1.46	1.32	.59	.43	1.68	-0.63	0.16	0.87	2.10
62	I feel embarrassed in front of other people too much	1.54	1.35	.67	.48	2.45	-0.62	0.11	0.67	1.64
35	Open up and tell my feelings to another person	1.70	1.40	.44	.33	0,95	-1.21	-0.02	0.83	2.26
3	Join in on groups	1.93	1.39	.62	.45	2.00	-1.00	-0.32	0.29	1.42
18	Tell personal things to other people	2.05	1.28	.52	.39	1.13	-1.89	-0.58	0.29	2.03
	Total	12.14	7.09	*H* = .42						α = .83

(continued)

		Descriptive and Mokken statistics				GRM parameters				
Nr	Item content	*M*	*SD*	r_{it}	H_i	*a*	b_1	b_2	b_3	b_4
	Nonassertive (HI)									
19	Be firm when I need to be	1.58	1.29	.55	.41	1.42	-0.92	-0.01	0.83	2.35
39	Be self-confident when I am with other people	1.61	1.31	.51	.38	1.25	-1.06	-0.01	0.92	2.25
12	Be another person's boss	1.69	1.32	.54	.41	1.34	-1.07	-0.15	0.80	2.06
9	Be assertive with another person	1.89	1.31	.73	.53	2.83	-0.99	-0.30	0.39	1.38
6	Tell a person to stop bothering me	1.94	1.36	.60	.44	1.67	-1.12	-0.40	0.30	1.61
8	Confront people with problems that come up	1.95	1.27	.63	.47	1.94	-1.35	-0.34	0.38	1.58
13	Be aggressive toward other people when the situation calls for it	2.00	1.40	.53	.40	1.31	-1.33	-0.49	0.34	1.56
5	Let other people know what I want	2.06	1.28	.66	.49	2.14	-1.26	-0.49	0.20	1.51
	Total	14.72	7.40	H = .44						α = .85
	Exploitable (JK)									
61	I let other people take advantage of me too much	1.06	1.17	.50	.39	1.20	-0.27	0.82	1.81	3.25
53	I am too gullible	1.25	1.22	.39	.30	0.79	-0.88	0.75	2.08	4.01
25	Argue with another person	1.42	1.20	.61	.45	1.89	-0.73	0.12	0.99	2.45
42	I am too easily persuaded by other people	1.53	1.24	.53	.39	1.25	-1.01	0.01	1.08	2.78
34	Feel angry at other people	1.90	1.29	.59	.43	2.02	-1.13	-0.38	0.33	1.70
10	Let other people know when I am angry	2.05	1.38	.64	.46	2.32	-1.07	-0.43	0.14	1.23
2	Say 'no' to other people	2.39	1.21	.62	.46	1.78	-1.88	-0.89	-0.17	1.26
38	Be assertive without worrying about hurting the other person's feelings	2.48	1.28	.53	.40	1.50	-1.93	-1.08	-0.32	1.04
	Total	14.08	6.73	H = .41						α = .83

(continued)

		Descriptive and Mokken statistics				GRM parameters				
Nr	Item content	*M*	*SD*	r_{it}	H_i	*a*	b_1	b_2	b_3	b_4
	Overly Nurturant (LM)									
49	I trust other people too much	1.09	1.17	.34	.27	0,71	-0.41	0.94	2.67	5.16
54	I am overly generous to other people	1.31	1.22	.53	.39	1.35	-0.61	0.29	1.36	2.82
63	I am affected by another person's misery too much	1.56	1.26	.52	.38	1.43	-0.99	0.07	0.93	2.33
51	I put other people's needs before my own too much	1.80	1.33	.69	.49	2.84	-0.86	-0.20	0.42	1.46
28	Have loving and angry feelings towards the same person	2.00	1.37	.48	.35	1.16	-1.58	-0.46	0.30	1.77
37	Attend to my own welfare when somebody else is needy	2.08	1.28	.59	.42	1.91	-1.41	-0.55	0.18	1.47
46	I try to please other people too much	2.11	1.27	.61	.44	1.80	-1.46	-0.59	0.13	1.53
21	Set limits to o ther people	2.28	1.25	.59	.43	1.62	-1.79	-0.81	-0.04	1.40
	Total	14.23	6.80	H = .40						α = .82
	Intrusive (NO)									
48	I want to be noticed too much	.94	1.13	.46	.28	0.95	-0.02	1.05	2.39	4.42
47	I clown around too much	1.07	1.23	.37	.23	0.70	-0.20	1.07	2.47	4.81
4	Keep things private from other people	1.21	1.22	.44	.26	1.21	-0.50	0.54	1.46	3.13
30	Stay out of other people's business	1.23	1.12	.37	.23	0.68	-1.12	0.70	2.70	5.90
26	Spend time alone	1.24	1.34	.19	.12	0.38	-0.93	1.55	3.40	6.45
58	I tell personal things to other people too much	1.24	1.25	.59	.34	3.77	-0.32	0.35	0.88	1.79
43	I open up to people too much	1.32	1.30	.54	.32	2.85	-0.33	0.24	0.80	1.90
41	I feel too responsible for solving other people's problems	1.69	1.29	.22	.15	0.37	-3.22	-0.25	2.09	6.58
	Total	9.94	5.63	H = .24						α = .70

Note: r_{it} = item-test correlation, under each subscale total M, SD, H_g value and Cronbach's α are given. H_i = item discrimination coefficient and H_g = scale discrimination coefficient (for nonparametric IRT scaling), a = discrimination parameter (for parametric IRT scaling), b_m = location parameter; the point at the latent trait continuum where there is a 50% chance of scoring in category m.

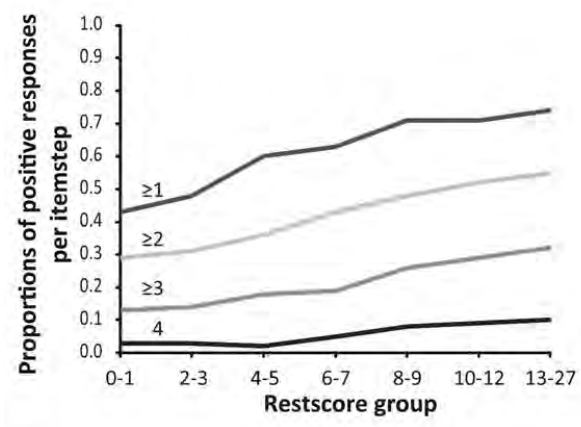

Figure 3.1: The item step response function for IIP-64 item 44 'I am too independent' of the Domineering subscale

Furthermore, results showed that items with high H_i values had increasing item step response functions and items with low H_i values had relatively flat item step response functions. To illustrate this, Figure 3.1 shows a relatively flat item step response function of item 44 with low H_{44} = .16 ("I am too independent", from the PA subscale). Persons with a relatively high rest score on PA, that is, persons that are assumed to show interpersonal problems related to dominant behavior, only have a slightly higher probability of choosing a particular response category as compared to persons with a low score on the PA (e.g., for person with rest score 13-27 the probability of choosing at least category 3 'quite a bit' equals .30, whereas the probability of person with rest score 2-3 equals .12). Thus, the item is unable to differentiate between persons with low, average, or high PA scores, and thus this item adds no information about a person's perceived dominant behavior to the PA subscale. We examined the subscales in more detail using the GRM.

3.3.2 Parametric IRT Analysis

In the last five columns of Table 3.1 the discrimination parameters and the location parameters estimated under the graded response model are given. We will not discuss these parameters on an individual level, but discuss the scale and item information functions that are based on these item parameters. Figure 3.2 shows the scale information curves (SIC), for the eight IIP-64 subscales. The curves depict the amount of information as a function of the estimated θ. On the x-axis, the estimated latent trait values are given in standard form. The

mean of the latent trait (θ = 0) reflects the mean of this specific clinical population. Information is inversely related to the conditional standard error of measurement, *SE(*θ), and thus the higher the information the higher the measurement precision. These graphs allow us to determine measurement precision for different ranges of θ. To assist with the interpretation of the graphs (see also Reise and Haviland, 2005), we drew three horizontal lines in Figure 3.2.

The lower line corresponds (approximately) to a reliability coefficient of .70 (*SE* = 0.548), the middle line to a reliability coefficient of .80 (*SE* = 0.447), and the upper line to a reliability coefficient of .90 (*SE* = .316). According to the guidelines for the interpretation of the reliability coefficient provided by Nunnally and Bernstein (1994) a value of .70 is sufficient for early stages of research, but basic research requires a reliability coefficient of .80 or higher, and when important decisions are to be made with test scores, a reliability coefficient of .90 is the minimum. Although the IIP-64 is used as a screening tool, and will always be used together with additional information from observation and intake interviews, it is interesting to consider the varying levels of measurement precision.

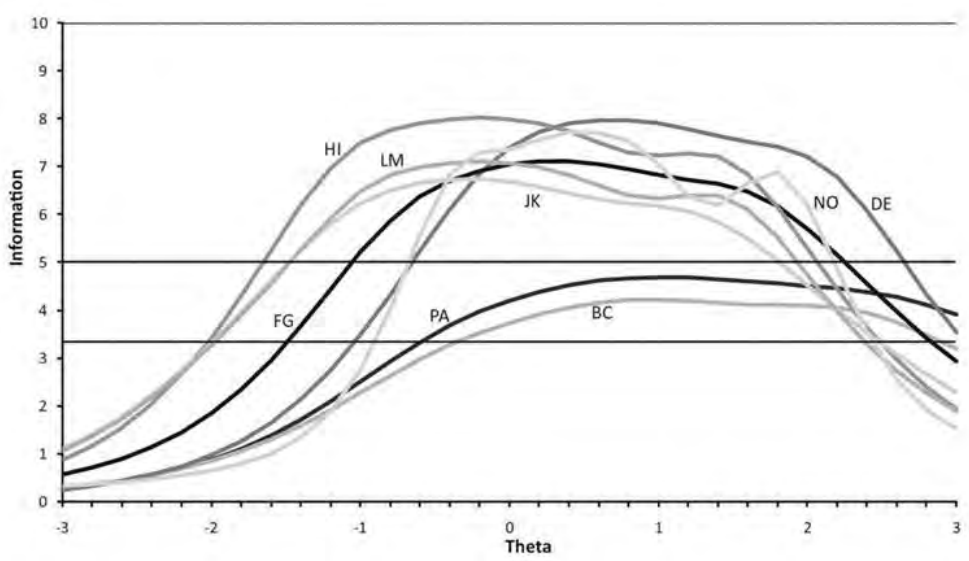

Figure 3.2: The scale information curves of the eight IIP-64 subscales. The lower, middle and upper dashed horizontal lines show the amount of information corresponding to a reliability of .70, .80, and .90 respectively. PA = Domineering, BC = Vindictive, DE = Cold, FG = Socially Avoidant, HI = Nonassertive, JK = Exploitable, LM = Overly Nurturant, NO = Intrusive

Table 3.2

Most Reliable Latent Trait Ranges and Descriptions of the Best and Worst Discriminating Items for Each Subscale.

		Content	
Subscale	Reliability > .80	Most relevant	Irrelevant
Domineering (PA)	-	Try to control (50), change (52), manipulate (57), and argue with others (59)	Too independent (44), difficult to take instructions (31) and understand others point of view (17)
Vindictive (BC)	-	Difficult to trust others (1), suspicious (56), get revenge (64)	Difficult to really care (24), be supportive (22) and put others needs before own (29), fight too much (40)
Cold (DE)	$-0.6 \leq \theta \leq 2.6$	Difficult to feel close (23), feel love (20) and show affection (15)	Difficult to forgive others (36)
Socially avoidant (FG)	$-1.0 \leq \theta \leq 2.2$	Difficult to join groups (3), socialize (14) and ask others to get together with me (33), easily embarrassed (62)	-
Nonassertive (HI)	$-1.6 \leq \theta \leq 2.0$	Difficult to be assertive (9) and let others know what I want (5)	-
Exploitable (JK)	$-1.4 \leq \theta \leq 1.8$	Difficult to argue (25), say 'no' (2), feel angry towards others (34) and let others know when I am angry (10)	Too gullible (53)
Overly nurturant (LM)	$-1.4 \leq \theta \leq 1,8$	Put other people's needs before own (51)	Too trusting (49)
Intrusive (NO)	$-0,6 \leq \theta \leq 2.0$	Tell personal things (58) and open up too much (43)	Wanting to be noticed (48), clowning around (47), feel responsible for others (41), difficult to spent time alone (26) and keep things private from others (4)

First, for all IIP-64 subscales $I \leq 8$, which corresponds to a reliability equal or smaller than 0.87 ($SE \geq .353$). For subscales PA and BC, we found $I \leq 5$ and a reliability smaller than 0.79 ($SE \geq 0.462$). The height of the discrimination parameters and H-values correspond with the amount of measurement precision given by the information curves, only the NO scale seems to have higher measurement precision than expected on the basis of the discrimination parameters.

Second, inspecting the SIC's we note that scale information varies across θ. For subscales PA, BC, DE, and NO, scale information is relatively high at the high θ levels and low at the

low levels. For $\theta < -1$ reliability drops below the .70. Because the item location parameters are located within the higher regions of the latent trait (see item location parameters bi in Table 3.1), measurement precision of these subscales is only sufficient for higher levels of θ, which means that persons at the lower levels of the latent trait are poorly measured because low scores on these scales are unreliable. Using the same line of reasoning we observe that for all IIP-64 subscales for the extreme θ-values (roughly $\theta < -2$ and $\theta > 2$) measurement precision is low. One may argue that low measurement precision in the lower latent trait values is to be expected because these scale are constructed so that they are sensitive to patients scoring high on, say, domineering or cold behavior.

A third interesting finding is that the items differed substantially in how much they add to the measurement precision of a scale. To illustrate this, consider the item information curves (IIC) for each IIP-64 subscale in Figure 3.3. Scale information is the summed item information across the eight items per subscale. For every subscale items can be identified that add little or no information to the total information. Subscale NO (intrusive) provides the most extreme example of the varying contribution of the IIC's to the SIC. Although most items have low *a*-values, resulting in a low overall *H*-value, measurement precision is reasonable due to two items with high *a*-values and, thus, large item information (item 43 'I open up to people too much' and item 58 'I tell personal things to other people too much'). Thus, although a person has to answer 8 items with respect to intrusiveness, there are only two items that are relevant, and it is unclear what the other items measure. Items like "I clown around too much" and "I feel too responsible for solving other people's problems" should not be interpreted as measuring a similar concept as an item like "I tell personal things to other people too much". With respect to the understanding and interpretation of the IIP-64 subscales for future clinical and research applications, Table 3.2 summarizes for each IIP-64 subscale the ranges of θ where the reliability is larger than .80, and item content that adds relevant information to the scale and item content that does not.

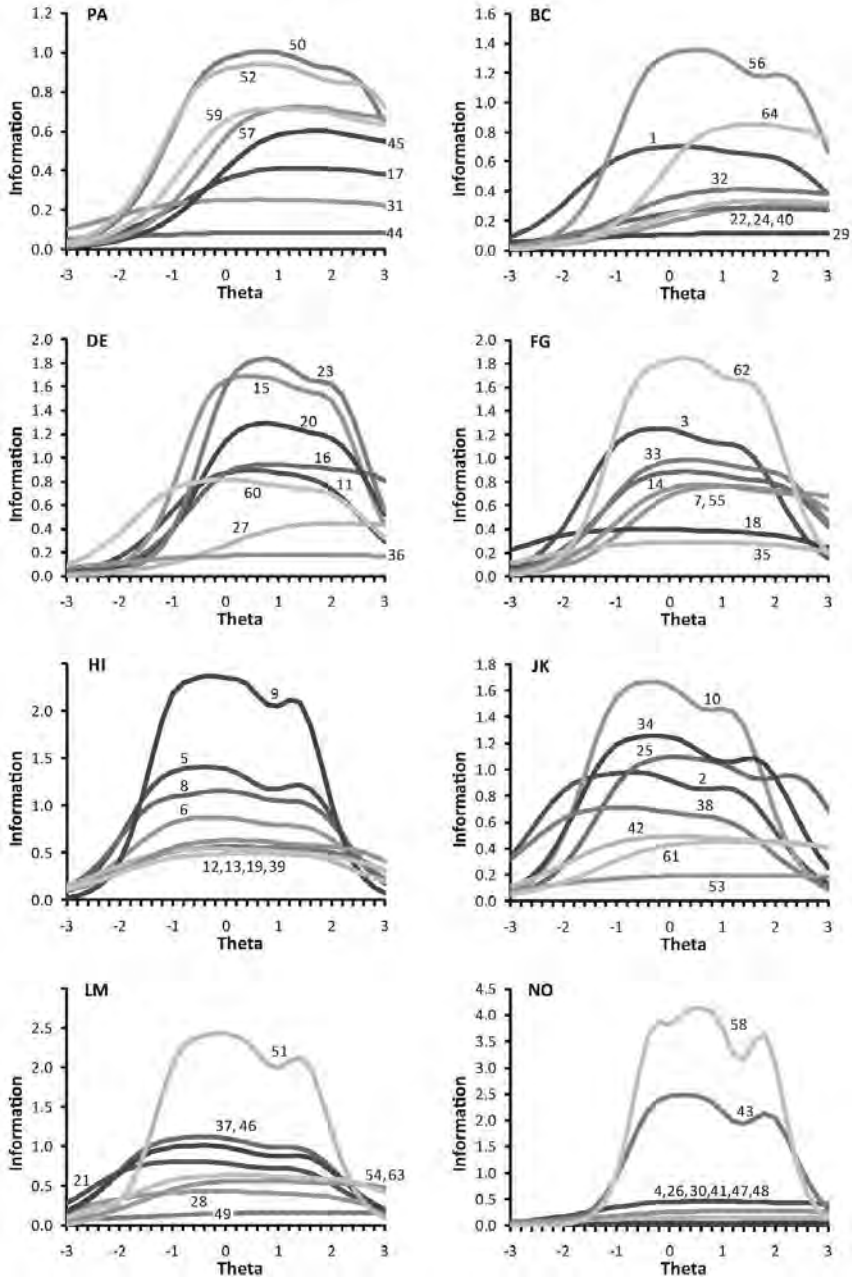

Figure 3.3: The item information curves for the IIP-64 items per subscale, which show the amount of information as a function of the latent trait each item contributes to the subscale. Note that the Y-axis are variable for better in-scale comparison of items, however there are large differences in amount of information across subscales.

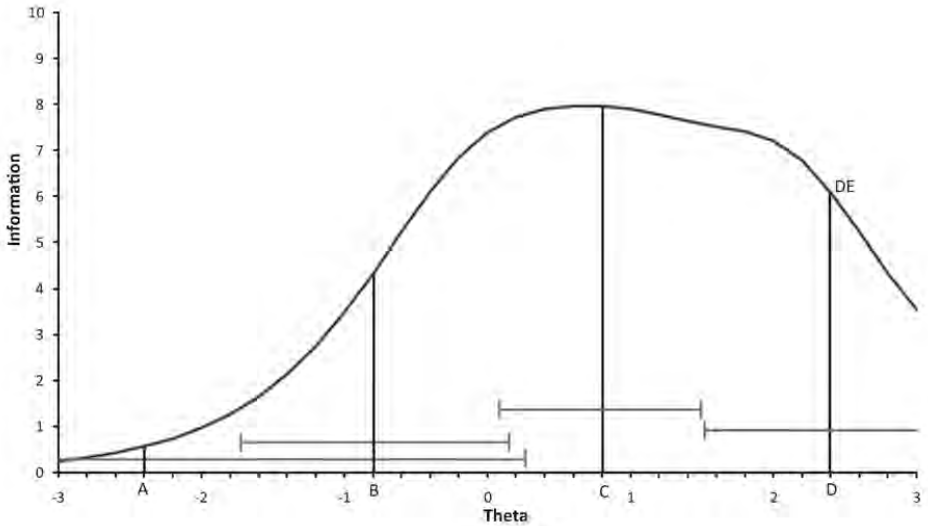

Figure 3.4: The scale information curve for the cold (DE) subscale with for persons A, B, C, and D their latent trait values $\hat{\theta}_A$ = -2.4, $\hat{\theta}_B$ = -0.8, $\hat{\theta}_C$ = -0.8 and $\hat{\theta}_D$ = 2.4 and corresponding 95% confidence intervals

To illustrate the consequences of these results for the interpretation of the IIP-64 subscale scores in practice, Figure 3.4 shows the scale information curve for the cold subscale (DE). Assume that there are four persons A, B, C, and D with varying severity of interpersonal distress related to the cold domain θ_A = -2.4, θ_B = -0.8, θ_C = 0.8 and θ_D = 2.4, respectively. Figure 3.4 depicts the 95% confidence intervals for these four persons based on the different standard errors for person A *SE* = 1.334, for person B *SE* = 0.481, for person C *SE* = 0.354, and for person D *SE* = 0.405. Persons A and B are located in the lower ranges of the DE latent trait scale. Confidence intervals are broad, especially for person A, and the only conclusion we can draw is that this person experience few to an average amount of interpersonal problems. Furthermore, note that although the difference between persons A and B is 1.6 points on the latent trait scale the 95% confidence intervals overlap. Although their scores differ, we cannot make a reliable distinction between persons A and B. However, because measurement precision is higher in the higher ranges of the latent trait scale, for person C and D with the same difference in latent trait values of 1.6 points, the 95% confidence intervals do not overlap. We can reliably state that person D has more interpersonal distress related to coldness as compared to person C.

3.3.3 Improving the Scales

To obtain better PA, BC, and NO scales we removed unscalable items and determined the scale quality of the remaining items. For the PA scale we found that removing the three items with $H < .30$ (items 17, 31, and 44) resulted in a scale with $H = .40$ and no decrease in information ($I \leq 5.419$). For the BC subscale, removing items 22, 24, 29, 32, and 40 resulted in a scale with $H = .51$ and $I \leq 6.796$. For the NO subscale, removing the two best items did not result in a new scale, the remaining items formed a scale with $H < .30$. The only items that resulted in a scale with $H \geq .30$ were the items 43 and 58. For the other subscales DE, FG, HI, JK, and LM removing the item with lowest H_i-value increased both total H and total information[2]. However, removing more than one item resulted in an increase of H but a decrease of total information. In other words, reducing the length of the subscales increased the scalability of the items at the expense of total information and thus measurement precision[3].

3.4 Discussion

The aim of the present study was to contribute to the existing literature of the IIP-64 by investigating the psychometric quality of the subscales. In particular the abilities of the items and subscales to differentiate between individuals with different severity of specific interpersonal problems, and the measurement precision for different individual subscale scores. We analyzed a large sample of clinical outpatients using nonparametric and parametric IRT approaches.

---

[2] For the DE subscale removing item 36 (with $H_i = .26$) increased H to .46 and $I \leq 8.848$, for the FG scale removing item 35 ($H_i = .33$) increased H to .45 and $I \leq 7.990$, for the HI subscale removing item 39 increased H to .46 and $I \leq 8.467$, for the JK subscale removing item 53 ($Hi = .30$) increased H to .45 and $I \leq 7.693$, and for the LM subscale removing item 49 increased H to .44 and $I \leq 8.041$.

[3] Interesting in this respect is that when we considered the subscales of the shortened version of the IIP, the IIP-32 (Horowitz et al., 2000) where each subscale consists of four items we found that for all but one subscale (PA) the overall H-values were larger as compared to the IIP-64, which reflects the narrower content of these scales. Highest total scale information however decreased for these subscales with an average of about $I = 2$ for subscales PA, DE, FG, HI, JK and LM, which is a substantial loss of information and measurement precision.

From our analysis it is clear that not all subscales consist of high-quality items. The subscales PA, BC, and NO consist of items that do not discriminate well between persons across ranges of interpersonal distress of each of these three domains. As a result, it is questionable whether the total scores on these scales can be used to order persons according to their (trait level) scores.

Another important finding is that for each subscale measurement precision varies across the range of the latent trait. In the extreme ranges of the latent trait measurement precision is low for all IIP-64 subscales. In addition, the four subscales PA, BC, DE, and NO provide low measurement precision for persons situated in the lower ranges of the latent trait scale. Persons cannot be ordered accurately in these ranges. Consequently, score differences in the lower ranges of the domineering, vindictive, cold, and intrusive interpersonal domains have little meaning. The observation that these scales provide an unequal amount of information across the latent trait scale is an often encountered phenomenon for clinical scales (Reise & Waller, 2009). The aim of measuring, for example, dominance is to detect persons high on dominance. As a result, items are selected with statements that indicate extreme dominant behavior which results in item location parameters for most items within a limited range of the latent trait and, consequently, the items provide most information within this limited range. Even so, we found that for the other four subscales FG, HI, JK and LM it was possible to differentiate between persons on both high and lower ranges.

The results of our empirical analyses can help researchers with decisions in using specific subscale scores, instead of only using total IIP scores (e.g., in effect studies such as Berghout, Zevalkink, & de Jong, 2011). Also, the results may help clinical researchers and practitioners to obtain a better understanding of the item content relevant to several domains of interpersonal problems their patients report. We suggest that future research may reconsider a revision of the PA, BC, DE and NO subscales of the IIP-64. Using both content information as well as information from IRT analyses, items can be selected that allow for constructs that are broad enough to have empirical validity and that provide acceptable measurement precision to distinguish patients on the different important constructs of interpersonal behavior. These scales should include items that tap the entire range of severity in order to differentiate between individuals and to be sensitive to change over time. We realize that it will not always be easy (perhaps sometimes impossible) to come up with subscales that are

both reliable and that measures constructs that are broad enough for prediction, but we think that IRT provides interesting tools to obtain better scales that measure the various domains of interpersonal problems.

Chapter 4
On the Factor Structure of the Beck Depression Inventory–II: G Is the Key

Abstract

The Beck Depression Inventory–II (BDI–II; Beck et al., 1996) is intended to measure severity of depression, and because items represent a broad range of depressive symptoms, some multidimensionality exists. In recent factor-analytic studies, there has been a debate about whether the BDI–II can be considered as one scale or whether subscales should be distinguished. In the present study, we applied a bifactor model to evaluate the extent to which scores reflect a single variable in a large sample of 1,530 clinical outpatients. We found that total scale score variation reflected some multidimensionality, but not enough to justify the scoring of subscales. We conclude that the BDI–II total scale score reflects a single construct and that reporting and interpreting subscale scores may result in misleading conclusions.

This chapter has been published as:

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2012, July 16). On the Factor Structure of the Beck Depression Inventory–II: G Is the Key. *Psychological Assessment*. Advance online publication.doi:10.1037/a0029228

4.1 Introduction

The Beck Depression Inventory–II (BDI–II; Beck et al., 1996) is used worldwide to assess the severity of depressive symptoms that correspond to the Diagnostic and Statistical Manual of Mental Disorders (4th ed., text rev.; DSM–IV–TR; American Psychiatric Association, 2000) criteria for major depressive disorder. The BDI–II is a revised version of the BDI–1A (Beck, Steer, Ball, & Ranieri, 1996). It consists of 21 items (see Table 4.2 in the results section for the item content). Each item consists of four statements, which are scored from 0 to 3. For example, Item 2 ("pessimism") has four response options ranging from 0 ("I am not discouraged about my future") to 3 ("I feel my future is hopeless and will only get worse").

The BDI–II is used to assess the severity of a patient's depression before clinical treatment for diagnostic purposes and with intervals during and after treatment to detect treatment progress or treatment stagnation. Mental health specialists use the BDI–II items to discover the depressive symptoms that are described in the DSM–IV–TR. Individual items may recover information about critical depressive symptoms, such as Items 2 and 9, which refer to suicidal ideation. A sum score consisting of the individual scores to the 21 items is used to estimate the overall severity of depresssion, and sometimes subscale scores are used to obtain information about specific domains of depressive severity, such as somatic or cognitive domains. Thus, depending on the focus of their investigation, clinicians sum item responses of the BDI–II to form one broad factor score or multiple narrow factor scores, and most often they do both. However, the interpretation of the scores with either approach may be problematic. To what extent can a sum score of all item responses be interpreted as representing a unidimensional factor of depression severity when at the same time subsets of these item responses can be interpreted as representing multiple factors of specific depressive symptoms? And to what extent is a subset of item responses specific for a particular subset of depressive symptoms if these items share common variance with the remaining BDI–II items?

Previous researchers have made different recommendations with respect to the underlying psychometric structure and dimensionality of the BDI–II and the usefulness of reporting subscale scores (e.g., Arnau, Meagher, Norris, & Bramson, 2001; Beck et al., 1996; Beck, Steer et al., 2002; Buckley, Parker, & Heggie, 2001; Dozois, Dobson, & Ahnberg, 1998;

Osman, Barrios, Gutierrez, Williams, & Bailey, 2008; Os-man et al., 1997; Steer, Ball, Ranieri, & Beck, 1999; Vanheule et al., 2008; Viljoen, Grant, Griffiths, & Woodward, 2003; Ward, 2006). In particular, different factor models have been fit to the BDI–II data, varying from a unidimensional factor model to different multidimensional two-and three-factor models, and various recent studies have reported different conclusions. In the present study, we investigated the dimensionality of the BDI–II, and the main question we addressed was whether the BDI–II total scale score variation primarily reflects (a) variation on a single construct, and thus the total scale score should unambiguously be interpreted as a unidimensional measure of depression, or (b) multiple nonignorable sources of variance, and consequently subscales for specific symptom groups need to be constructed.

4.1.1 Recent studies

Several researchers have addressed questions regarding the structural validity of the BDI–II in recent years. One important contribution was made by Ward (2006), who used a factor model that separated the role of a general (G) factor and the role of group factors. Ward compared two two-factor models (one with a Somatic-Affective and Cognitive factor and one with a Cognitive-Affective and Somatic factor) with a group factor model for three clinical and three nonclinical data sets from previously published factor-analytic studies of the BDI–II. In the group factor model, the 21 BDI–II items were direct indicators of a general factor, and they were also allowed to load on two group factors (a Somatic and Cognitive group factor). These group factors and the general factor were assumed to be orthogonal and thus uncorrelated. With this model, it was possible to distinguish variance that could be attributed to the general factor and variance that could be attributed to the group factors. The group factor model yielded a superior fit in the six samples studied as compared to the other models. There was a strong general factor, explaining on average 76% of the total common variance. The contributions of the group factors were relatively small (6%–14%). Ward concluded that because most of the total score variance in each subscale was due to the general factor, subscale scores based on previous factor models (a) are difficult to interpret and (b) have limited reliability and discriminant validity. The group factor model can be conceived as a bifactor model (Gibbons & Hedeker, 1992; Holzinger & Swineford, 1937;

Reise et al., 2007) and was also used in the present study. In the remainder of this article, we use the term *bifactor model*[1] instead of *group factor model*.

A second important study was done by Vanheule et al. (2008). They used a clinical sample of 404 outpatients. Comparing the fit of different factor models, they concluded that two models had a better fit to clinical data: (a) a two-factor model proposed by Dozois et al. (1998) with a Cognitive-Affective and Somatic-Vegetative factor and (b) a three-factor model proposed by Beck et al. (2002) with three factors labeled Cognitive, Somatic, and Affective. Both models are correlated-traits factor models with first-order factors. Vanheule et al. (2008) also analyzed the data using the bifactor model that was used by Ward (2006) and found that the model fulfilled most criteria for good fit, Satorra–Bentler χ^2 (174) = 357.02, comparative fit index (CFI) = .939, root-mean-square error of approximation (RMSEA) = .053, but they also observed that for a clinical sample, Items 8 and 18 had low or negative factor loadings on the group factors. Furthermore, they concluded that a unidimensional model did not show a good fit to clinical BDI–II data, Satorra–Bentler χ^2 (189) = 587.12, CFI = .868, RMSEA = .075. They favored the use of subscale scores, stating: '*We believe that the inclusion of a G factor which loads on all items is problematic: It is difficult to interpret what this G factor measures, or to implement it in research and practice. Moreover, its inclusion implies that the other subscales are not unidimensional and difficult to interpret (e.g., what the cognitive factor means if all variance of common depressive severity is extracted from it).*' (p. 180) In the current study, however, we discuss and show that ignoring the general factor leads to problematic measurement.

In another comprehensive review of factor models Osman et al. (2008) compared four correlated-traits factor models and one bifactor model in a nonclinical sample of 414 adolescents. They found the best model fit, χ^2 (168) = 414, CFI = .945, RMSEA = .043, for a bifactor model with a general factor and a Cognitive-Affective and Somatic group factor from the student sample of the Beck et al. (1996) study. The general factor in the bifactor model accounted for 68% of the common variance, whereas the contributions of the group

[1] Holzinger and Swineford (1937) originally named this model the bifactor model. Other labels are "group-factor model", "general-specific model" and "nested factor model". Note that a bifactor model is different from a correlated-traits factor model or higher-order model (see Gustafsson & Åberg-Bengtsson, 2009).

factors were 11% and 21%, respectively. They concluded that '*most of the BDI–II items are related either moderately or highly to a general factor*' (p. 98). Recently, Quilty et al. (2010) investigated different factor structures of the BDI–II in a sample of 425 outpatients with a major depressive disorder. They recommended using the bifactor model proposed by Ward (2006). They also evaluated this model by means of factor associations with an external, interviewer-rated measure of depression severity as assessed by a clinical interview. They found that the general factor taps both the presence of negative affect and the lack of positive affect. This adds to the evidence for the interpretation of the general factor. Their results thus supported the fit of a bifactor model and the use of the total scale score. They concluded that (a) correlations between subscales were high in models without a general factor where factors are allowed to correlate, (b) the bifactor model shows a good fit across multiple samples, and (c) the model retained good fit without correlated errors where other models did not.

Finally, Al-Turkait and Ohaeri (2010) compared the goodness of fit of correlated-traits two- and three-factor models with higher order and bifactor models of the BDI–II in a sample of 624 Arab college students. The bifactor model fitted best (CFI = .91, Tucker–Lewis index [TLI] = 0.89, RMSEA = .042) as compared to the other models. The regression weights of the bifactor models showed that the variance related to the group factors was mostly accounted for by the general depression factor.

4.1.2 Subscales and the General Factor

The studies cited above show that there is no clear consensus on the best fitting model, which results in different recommendations regarding how to best score and interpret BDI–II results. A reason for the lack of consensus may be that to capture the complex construct of depression, the BDI–II consists of items that represent a broad range of depression criteria. As a result, the BDI–II is not clearly one- or multidimensional. The items measure the common construct depression and at the same time contain item clusters that measure one specific aspect of depression (e.g., a somatic aspect). For example, Items 15 ("loss of energy") and 20 ("tiredness or fatigue") measure something that is common to the construct of depression. At the same time, these two items form a cluster that shares unique variance (namely, the explicit somatic content of these items) compared to the other items. A more general question that runs through all the BDI–II studies conducted thus far is, how should we analyze a clinical instrument that measures one thing (depression) and at the same time

measures the same thing in a slightly different way (somatic, cognitive, or affective elements of depression)?

In recent years, several researchers have addressed this more general topic of measuring psychological constructs at different levels of the construct hierarchy. Results in the area of cognitive abilities and intelligence (e.g., Carroll, 1995; Gignac, 2005, 2006; Watkins, 2010), quality of life, psychopathology, and personality measurement (e.g., Chen et al., 2006; Gignac, 2007; Reise et al., 2007) suggest that psychological data seldom if ever have a clean dimensionality structure. Gustafsson and Åberg-Bengtsson (2010) provided a historical overview and empirical examples of different types of hierarchical models. They discussed the idea that instrument homogeneity is neither a necessary nor sufficient principle for achieving instruments that are practically and theoretically useful, and that to avoid the problem of under-representing a complex construct, such as depression, measures are constructed with heterogeneous content. Reise et al. (2010; for an overview, see Brunner et al., 2012; Reise, 2012) made an important contribution to the discussion about whether we should provide total or subscale scores when using clinical questionnaires with heterogeneous content. They argued that the often chosen solution of reporting both total scale score and subscale scores is problematic: First, in a unidimensional model the effect of a specific group of items (or symptoms) within the broader construct can be overlooked. For example, variations in total scale score or correlations with external variables can be attributed to the broader construct, whereas they are in fact strongly influenced by the effect of a specific group of items within that construct. Second, in a multidimensional model, multicollinearity can interfere with the ability to judge the unique contribution of each subscale. Third, often psychologists use subscales because of the actual or assumed different correlates with external variables. “However, any two items that are not perfectly correlated potentially have different correlations with external variables. Yet it would be silly to argue that one should investigate correlations for each item separately” (Reise et al., 2010, p. 554). Creating short subscales often results in scale scores that are less reliable than the original total scale score. And fourth, because in clinical measures subscales often reflect variation on both a general construct (depression) and more specific constructs (e.g., somatic elements of depression), subscale scores may appear reliable, not due the unique variance that is explained, but due to the general variance that is also measured by the subscale.

Reise et al. (2010, 2007) recommended using a bifactor model to analyze clinical questionnaires with heterogeneous content. The bifactor model can be used to evaluate the extent to which scores reflect a single variable even when the data are multidimensional. A bifactor model can complement traditional dimensionality investigation by evaluating whether item response variance is due to a general construct versus group factors[2]. This evaluation is of specific interest in case of the BDI–II, because previous factor-analytic approaches of various factor models have yielded little agreement as to which model should be applied to BDI–II data. Some recent studies indeed favored the bifactor model to other factor models to describe BDI–II data (Al-Turkait & Ohaeri, 2010; Osman et al., 2008; Quilty et al., 2010; Ward, 2006), whereas other studies had conceptual problems using the bifactor model (Vanheule et al., 2008).

In the present study, we investigated whether BDI–II scale score variation is mainly due to a single general factor or to multiple group factors in a large sample of clinical outpatients. We replicated analyses of three-factor models that were found to show the best fit to clinical BDI–II data in two of the recent articles discussed above: the Vanheule et al. (2008) study and the Quilty et al. (2010) study. We used a one-factor model, a two-factor model (Dozois et al., 1998), a three-factor model (Beck et al., 2002), and a bifactor model (Ward, 2006). In addition, we compared the two-and three-factor models, with their corresponding bifactor models.

---

[2] An alternative model that is often used to distinguish between item response variance due to general and specific factors (or higher- and lower-order factors) is the higher-order model with a Schmid-Leiman transformation (Schmid & Leiman, 1957). In a higher-order model the relations between a higher-order factor and item responses are mediated by the lower-order factor, which imposes a proportionality constraint on the variance ratios of general and specific effects in item responses while the bifactor model does not. A Schmid-Leiman exploratory bifactor analysis can be used prior to fitting confirmatory bifactor models, for example, to identify item cross-loadings. The confirmatory bifactor model provides more insight into the relationship between general and specific factors in explaining item response variance (Brunner et al., in press; Gignac, 2008; Reise, in press; Reise et al., 2007; for a formal comparison of these two models, see Yung, Thissen & McLeod, 1999).

4.2 Method

4.2.1 Participants

The sample consisted of 1,530 outpatients (61.5% female and 38.5% male). Mean age was 35.1 years (SD = 10.1) for the entire sample, 33.4 years (SD = 9.5) for women, and 37.8 years (SD = 10.3) for men—a moderate gender difference, $t(1524) = 8.493$, $p = .001$, Cohen's $d = 0.44$. Of the outpatients, 87.5% reported Dutch as their dominant culture when asked in which culture they were raised, and 62.6% had a bachelor's or master's degree. Data were obtained as part of a psychological screening assessment procedure for persons applying for treatment in the period between the years 2002 and 2010 at a community mental health clinic specializing in ambulatory psychoanalytic treatment. The included patients gave informed consent and completed the BDI–II[3]. Psychiatric diagnoses were assessed (by at least one psychiatrist and two other registered mental health specialists) in a consensus meeting, and patients were classified according to the DSM–IV– TR. For 93.5% of the 1,530 participants, DSM–IV–TR Axis I classifications were available. Most frequent classifications of clinical syndromes were mood disorders (41.8%), anxiety disorders (19.0%), adjustment disorders (19.0%), and substance-related disorders (8.6%). In addition, on Axis I of the DSM–IV–TR, additional problematic conditions were classified that could not be classified as clinical syndromes but were serious enough to warrant independent clinical attention. Most frequent classifications of additional problematic conditions were partner-relational problems (27.1%), identity problems (26.6%), and phase-of-life problems (11.4%). Axis II classifications were not assessed systematically.

[3] Participants also completed three other clinical questionnaires: The Symptom Checklist-90 (Arrindell & Ettema, 1986; Derogatis, 1983), the State–Trait Anxiety Inventory (Spielberger, 1998; Van der Ploeg, 2000) and the Inventory of Interpersonal Problems 64 (Horowitz, Alden, Wiggins, & Pincus, 2000).

4.2.2 Analysis

Table 4.1 gives an overview of the BDI–II factor models we used:

- Model A: the one-factor model with all 21 items loading on one factor;
- Model B: the two-factor model used by Dozois et al. (1998), with 10 items loading on the Cognitive-Affective factor (Items 1–3, 5–9, 13, 14) and 11 items loading on a Somatic-Vegetative factor (Items 4, 10 –12, 15–21);
- Model C: a three-factor model from the Beck et al. (2002) study, with seven items loading on a Cognitive factor (Items 3, 5– 8, 13, 14), nine items loading on a Somatic factor (Items 10, 11, 15–21), and five items loading on an Affective factor (Items 1, 2, 4, 9, 12);
- Model D: a bifactor model with a general factor and two group factors based on the two-factor model (Model B);
- Model E: a bifactor model with a general factor and three group factors based on the three-factor model (Model C); and
- Model F: the bifactor model from the Ward (2006) study, with a general factor, a five-item Somatic group factor (Items 15, 16, 18 –20), and an eight-item Cognitive group factor (Items 2, 3, 5–9, 14).

As discussed above, the bifactor model provides a valuable tool to investigate whether item variance is due to the general factor (depression) or to specific factors (such as a somatic or cognitive symptoms of depression). In a bifactor model, each item loads on a general factor and is also allowed to load on one of the two or more orthogonal group factors that are specified. There is a general factor that explains the item intercorrelations, and in addition, there are group factors that explain the item intercorrelations that attempt to capture the residual variation due to secondary dimensions. We compared the results from the one-factor and correlated-traits two- and three-factor models with those from the bifactor models (for a similar approach, see Reise et al., 2007).

Table 4.1

Overview of the BDI-II Factor Models that were used in the Current Study.

Factor	Short form	Items
Unidimensional		
Model A		1-21
Correlated-traits		
Model B		
Cognitive-Affective	CA	1-3, 5-9, 13, 14
Somatic-Vegetative	SV	4, 10-12, 15-21
Model C		
Cognitive	C	3, 5-8, 13, 14
Somatic	S	10, 11, 15-21
Affective	A	1, 2, 4, 9, 12
Bifactor		
Model D		
General	G	1-21
Cognitive-Affective	g_{CA}	1-3, 5-9, 13, 14
Somatic-Vegetative	g_{SV}	4, 10-12, 15-21
Model E		
General	G	1-21
Cognitive	g_{C}	3, 5-8, 13, 14
Somatic	g_{S}	10, 11, 15-21
Affective	g_{A}	1, 2, 4, 9, 12
Model F		
General	G	1-21
Somatic	g_{S}	2, 3, 5-9, 14
Cognitive	g_{C}	15, 16, 18-20

Note: Model B is the two-factor model from the Dozois et al. (1998) study, model C is the three-factor model from the Beck et al. (2002) study and model F is the bifactor model from the Ward (2006) study.

First, we investigated which model demonstrated the best fit to the data. Second, we compared the factor loadings for the general factor in the bifactor model with those for the one-factor model. Considerably lower factor loadings for the general factor indicate that variance in item responses is influenced by the group factors and thus that the data are not

unidimensional. Third, we compared the item loadings from the group factors in the bifactor model with those from the two- and three-factor models. Discrepancy between the loadings indicates the degree to which item variance in the correlated-traits factor models remains specific in the bifactor models, after the common variance is accounted for. Fourth, we compared the factor loadings of the general factor with those of the group factor within the bifactor models. The difference indicates the degree to which items reflect the general factor or a specific group factor in a bifactor model. Fifth, as an index of unidimensionality, we calculated the percent of explained common variance (ECV) that was attributable to the general factor and to group factors (Bentler, 2009; Reise et al., 2010; Ten Berge & Socˇan, 2004). For each factor the ECV is the sum of squared factor loadings for that factor divided by the sum of all squared factor loadings (the common variance) for the model. Sixth, we compared the reliability of scale scores for all factors. For the one-factor model and correlated-traits factor models (Models A, B, and C), we calculated omega (ω), which is an index for the proportion of variance accounted for by a factor relative to the total observed score variance (where $0 \leq \omega \leq 1$ and $\omega = 1$ indicates that the sum score measures the target construct with perfect accuracy). Because in a bifactor model each item response is assumed to be influenced by both general depressive symptomatology and specific depressive symptoms, we calculated scale score reliability for the factors in the bifactor models (Models D, E, and F) using both ω and omega-hierarchical (ω_h). ω_h indicates the proportion of variance in a scale score that is accounted for by what is specific for a subset of item responses to the total observed variance for these item responses (see Brunner et al., 2012; Reise, 2012; for a comparison with other reliability indices, see Zinbarg, Revelle, Yovel, & Li, 2005).

The confirmatory one-, two-, three-, and bifactor models were estimated with Mplus 4.1 (Muthén & Muthén, 2006). Because the observed variables in the models were categorical and Mardia's normalized estimate of multivariate kurtosis was indicative of nonnormality (z = 45,389, p = .000; Mardia, 1970), the mean and variance-adjusted weighted least squares estimation was used for all calibrations. We used the following fit indices and rules of thumb: the CFI, good fit if CFI ≥ .95 and acceptable fit if CFI is between .90 and .95; the TLI, good fit if TLI ≥ .90; and the RSMEA, good fit if RSMEA ≤ .06 and acceptable fit if RMSEA is between .06 and .08 (see Cook, Kallen, & Amtmann, 2009; Hu & Bentler, 1998; for a critical discussion, see Reise, 2012; Reise, Scheines, Widaman, & Haviland, 2012).

4.3 Results

4.3.1 Descriptive Statistics

Table 4.2 presents descriptive statistics for the BDI–II item and scale scores. Cronbach's α for the BDI–II total scale equaled .90 (95% CI [.89, .91]). The corrected item-total

Table 4.2
Descriptive Statistics for the BDI-II Items in a Sample of N = 1530 Clinical Outpatients.

Nr	Content	*M*	*SD*	%	r_{it}	Skewness	Kurtosis
1	Sadness	.85	.71	70	.63	.86	1.32
2	Pessimism	1.01	.86	71	.59	.68	-.06
3	Past failure	.94	.94	57	.53	.46	-1.05
4	Loss of pleasure	.98	.84	69	.61	.48	-.46
5	Guilty feelings	.87	.81	64	.51	.72	.06
6	Punishment feelings	.60	.99	33	.39	1.50	.89
7	Self-dislike	1.21	.91	75	.58	.18	-.87
8	Self-criticalness	1.19	.93	72	.53	.16	-1.00
9	Suicidal thoughts	.45	.57	42	.48	.98	.76
10	Crying	1.08	1.05	65	.42	.70	-.69
11	Agitation	.91	.75	71	.47	.80	.82
12	Loss of interest	.87	.86	62	.63	.90	.27
13	Indecisiveness	1.18	1.13	63	.54	.48	-1.17
14	Worthlessness	.92	.94	56	.60	.47	-1.08
15	Loss of energy	1.25	.84	79	.62	-.01	-.83
16	Changes in sleep	1.20	.99	72	.44	.42	-.84
17	Irritability	.90	.83	64	.50	.62	-.25
18	Changes in appetite	.75	.88	52	.46	1.04	.30
19	Concentration difficulty	1.21	.89	74	.61	.03	-1.01
20	Tiredness or fatigue	1.13	.90	73	.60	.40	-.63
21	Loss of interest in sex	.66	.88	43	.37	1.14	.25
	Total	20.13	10.80	α = .90			

Note: % = percentage of persons endorsing response options 1, 2, or 3, r_{it} = corrected item-test correlation. In the last row the total scale *M, SD*, and Cronbach's *α* are provided.

correlations ranged from .39 to .63 (lowest item-total correlations for Items 6 and 21 with r = .40). These descriptive statistics corresponded with the descriptive statistics for BDI–II items reported in other studies (e.g., Beck et al., 1996; Beck et al., 2002).

4.3.2 Confirmatory Factor Analysis

Table 4.3 displays the factor loadings, proportions of variance explained, and reliability indices for the factors and the fit indices for the different factor models. As described above, we analyzed the fit indices and factor loadings in six steps. First, in terms of model evaluation, fit indices demonstrated that the one-factor model had a good fit to the data according to the TLI, but not according to the CFI and the RMSEA. The correlated-traits factor models showed acceptable model fit according to CFI and RMSEA criteria and a good fit according to the TLI. The three bifactor models appeared to be equally well fitting models and demonstrated the best model fit of the models we tested. Second, the item loadings in the one-factor model were only slightly lower as compared to the loadings on the general factor in the bifactor models. On average the loadings differed .03. This indicates that the loadings for the one-factor model were not seriously distorted by multidimensionality. Third, the factor loadings in the correlated-traits factor models were high (with an average of .67 ranging from .47 to .80), suggesting that for these models the items discriminate well between persons. However, inspection of the bifactor results showed that the loadings on the group factors were much lower after controlling for the general factor (with an aver-age loading of .28 ranging from ₃.17 to .69). For the Cognitive group factors there were several items with loadings larger than.40. However, for the other group factors the factor loadings were low. Note that many of the factor loadings for the group factors in the bifactor models were smaller than .20. These solutions were empirically underidentified. In Model E, for example, there was only one item (Item 12) that uniquely loaded on the Affective group factor. This is not an interpretable factor. Fourth, for the bifactor models each item had a higher factor loading on the general factor than on the group factor, except for two items (Items 3 and 20).

Table 4.3

Item loadings, Proportions of Variance Explained, Reliability Indices and Fit Statistics for One, Two, Three and Bifactor CFA Models for the BDI-II.

Item	One and correlated-traits factor models						Bifactor models									
	Model A	Model B		Model C			Model D			Model E				Model F		
		CA	SV	C	S	A	G	g_{CA}	g_{SV}	G	g_C	g_S	g_A	G	g_S	g_C
1	.76	.80				.80	.77	.13		.80			.13	.79		
2	.68	.73				.73	.63	.34		.73			.17	.63		.32
3	.64	.69		.71			.48	.62		.56	.49			.50		.61
4	.71		.75			.75	.76		.02*	.74			-.17	.74		
5	.57	.61		.64			.47	.45		.50	.45			.48		.44
6	.49	.52		.54			.42	.32		.45	.28			.42		.32
7	.67	.71		.74			.56	.48		.60	.46			.57		.48
8	.61	.65		.67			.50	.47		.52	.48			.51		.46
9	.61	.64				.64	.60	.19		.65			.21	.59		.18
10	.49		.51		.52		.52		.03*	.49		.09		.51		
11	.56		.59		.61		.56		.19	.53		.26		.59		
12	.74		.78			.79	.77		.09	.79			-.48	.78		
13	.63	.67		.70			.62	.18		.63	.16			.66		
14	.72	.76		.80			.60	.53		.65	.46			.61		.51
15	.75		.79		.81		.67		.47	.65		.51		.69	.46	
16	.51		.54		.55		.47		.30	.45		.35		.49	.26	
17	.58		.61		.63		.57		.22	.54		.28		.61		
18	.52		.55		.56		.51		.18	.49		.23		.52	.13	
19	.69		.73		.74		.69		.22	.66		.27		.69	.20	
20	.73		.77		.79		.60		.69	.59		.67		.65	.64	
21	.45		.47		.48		.44		.18	.42		.21		.47		
Eigenvalue	8.36	4.65	4.70	3.33	3.71	2.77	7.32	1.63	1.00	7.60	1.13	1.15	.35	7.65	.75	1.51
ECV		50%	50%	34%	38%	28%	74%	16%	10%	74%	12%	11%	3%	77%	8%	15%
ω	.85	.90	.89	.86	.89	.86	.94	.90	.89	.94	.87	.86	.88	.94	.84	.88
ω_h							.83	.27	.12	.85	.34	.22	.00	.86	.20	.33
CFI	.796	.884		.903			.943			.929				.926		
TLI	.938	.965		.972			.983			.979				.978		
Free parameters	84	85		87			105			105				97		
RMSEA	.106	.079		.071			.055			.062				.063		

Note. Factor loadings were estimated under mean and variance-adjusted Weighted Least Squares estimation. G = general factor, CA = Cognitive-Affective, SV = Somatic-Vegetative, C = Cognitive, S = Somatic and A = Affective, g_x are group-factors for the bifactor models. Correlations between factors are $r_{CA\text{-}SV}$ = .78 for model B and $r_{C\text{-}S}$ = .69, $r_{S\text{-}A}$ = .82 and $r_{C\text{-}A}$ = .80 for model C; ECV = percentage of explained common variance; ω = omega reliability coefficient; ω_h = omega hierarchical; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation, * = non-significant factor loadings.

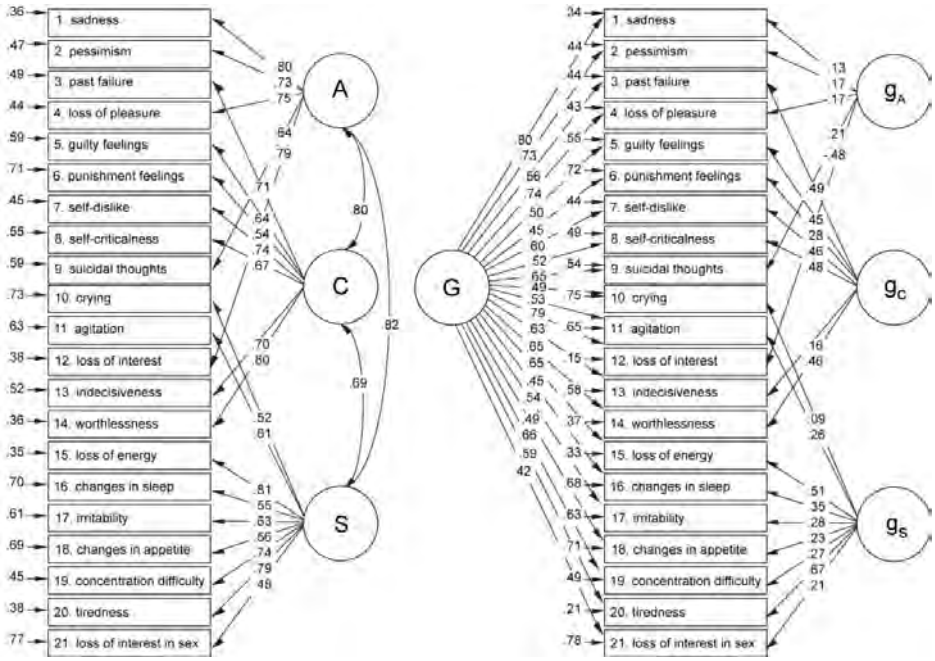

Figure 4.1: Standardized factor loadings, correlations between factors and measurement error terms for models C and E. Model C on the left side is the correlated-traits three factor model from the Beck et al. (2002) study, model E on the right side is the corresponding bifactor model with three group factors. C = Cognitive, S = Somatic, A = Affective, G = General factor, g_x are group-factors for the bifactor models.

To graphically illustrate the differences between the measurement models, we chose the correlated-traits factor Model C and the corresponding bifactor Model E, as they are relevant in demonstrating our findings and because these particular subscales are often used in clinical practice (see Figure 4.1). After controlling for the general factor, the factor loadings for the Affective group factor were very low, and for the Somatic group factor only two items had loadings larger than .40. Items 3, 5, 7, 8, and 14 from the Cognitive group factor with factor loadings larger than .40 formed the strongest cluster of items.

The following two steps demonstrate the consequences of these findings in terms of the common variance that was explained by the different factors in each of the models and the reliability of scale scores. The explained common variance of all models ranged from 44% to 49%. In the two-factor correlated-traits model (Model B), both the Cognitive-Affective and

Somatic-Vegetative factors explained 50% of the common variance. However, in the corresponding bifactor model (Model D), the general factor accounted for 74% of the common variance, and 16% and 10% were explained by the group factors, respectively

In sum, for the three bifactor models, 75% of the common variance was attributable to a single general factor. Finally, the reliability of the summed total scale score, based on the bifactor results, ranged from $\omega_h = .83$ to.86. This means that 83%– 86% of the variance of this summed score is attributable to the general factor. We recommend reporting an estimated reliability of approximately .85 as opposed to the somewhat misleading alphas, which generally are larger than .90. The estimated reliabilities for summed subscale scores after controlling for the general factor were at most $\omega_h = .34$. That is, at most 34% of the variance in the subscale scores is explained by the specific content of the subscale items beyond the variance that is already explained by the general factor. This supports the assump-tion that if scores of subscales for the BDI–II are used, their interpretation as precise indicators of unique constructs is limited because very little reliable variance exists beyond that due to the general factor.

4.4 Discussion

In this study, we evaluated the factor structure of the BDI–II with the main aim to investigate the extent to which item responses reflect one single or multiple variables. We compared the results from one-, two-, and three-factor models with those from three bifactor models in a large sample of clinical outpatients (the target population of the BDI–II). We observed that total scale score variation reflected multiple sources of variance due to clustered item content (especially a cluster of Items 3, 5, 7, 8 and 14 and a cluster of Items 15 and 20). However, differences in factor loadings between the unidimensional model and the general factor from the bifactor models were small, and the ECV of the general factor for all bifactor models was large (>74%). Reise et al. (2012) found that when the ECV for the general factor in a bifactor model is larger than 60%, the factor loading estimates for a unidimensional model are close to the true loadings on the general factor in the bifactor model. Consequently, we conclude that the presence of multidimensionality does not handicap our ability to interpret the BDI–II as one scale. In fact, on the basis of the current results, clustering of items into separate dimensions and consequently scoring of subscales can hardly be justified, because (a) only a

small number of items had factor loadings larger than .40 after the general factor was accounted for, and these items alone do not support the creation of subscales; (b) the general factor accounted for 74%–77% of the common variance in the three bifactor models, and the group factors only 3%–16%; and (c) in the bifactor models the reliabilities of the total scale scores ranged from .83 to.86, but for subscale scores as measuring a specific construct after controlling for the general factor, ω_h was .34 at best. The clinical relevance of our empirical findings is that there is more common variance to the BDI–II factors than unique variance. This implies that clinical practitioners should be careful when interpreting subscale scores, because these subscale scores are highly related to the general construct.

Our findings are in line with the results and conclusions of Al-Turkait and Ohaeri (2010), Osman et al. (2008), Quilty et al. (2010), and Ward (2006). These authors also concluded that most variance in each subscale score is common rather than specific. If a latent factor does not represent the common variance among a set of diverse items, then it is very difficult to interpret what that latent factor is measuring. The bifactor model can identify the factor that represents the common variance among a set of diverse items, in case of the BDI–II we interpret this factor as depression severity. Although Vanheule et al. (2008, p. 180) stated that "the inclusion of a G factor which loads on all items is problematic: It is difficult to interpret what this G factor measures," our results based on bifactor analyses lead us to the opposite conclusion, namely, that *not* including a general factor may lead to wrong interpretations.

It is important to realize that most psychological measures mix variation from different levels of the construct hierarchy and that ignoring a level can be problematic. Many psychological constructs operate at different levels of generality ranging from broadband (e.g., depression) to conceptually narrow constructs (e.g., somatic elements of depression; Brunner et al., 2012; Emmons, 1995; Reise, 2012). This is inherent to the nature of complex constructs in an area such as clinical psychology. When trying to capture the breadth of these constructs, researchers need to include indicators (items) with heterogeneous content, which places them "in the vexing position of trying to measure one thing while simultaneously measuring diverse aspects of this same thing" (Reise et al., 2010, p. 545). Consequently, clinical questionnaires often contain one broad common source of variance and multiple narrow sources of variance due to clustered item content (e.g., Brouwer et al., 2008; Gibbons, Rush, & Immekus, 2009; Meijer, de Vries, & van Bruggen, 2011; Simms, Grös, Watson, & O'Hara,

2008). The results of this study demonstrate that when researchers choose a measurement model without a general factor, scale scores for specific constructs are difficult to interpret (see also Brunner et al., 2012; Gustafsson & Åberg-Bengtsson, 2010; Reise, 2012). Researchers have also used the bifactor model in the cognitive domain to evaluate factor structure. For example, Gignac (2005, 2006) conceptualized the Wechsler Adult Intelligence Scale (revised and third editions) as measuring a general factor and group factors, and Watkins (2010) distinguished different levels of generality for the Wechsler Intelligence Scale for Children (4th ed.; WISC–IV) and found that the general factor was the predominate source of variation among WISC–IV subtests.

On the basis of the factor-analytic results from this study, we have made recommendations with respect to the structural validity of the BDI–II. However, factor analysis alone is insufficient for testing scale validity. For future validity research, our results imply that it is insufficient to relate subtest scores to external variables without taking the general factor into account. Relations of both broad and narrow dimensions to external criteria are critical when researchers investigate the validity of test score interpretations. Brunner et al. (2012) discussed that all constructs that are specified in a bifactor model can be linked to external variables. This provides the basis for future research regarding the validity of the subscale interpretations of the BDI–II. Researchers should investigate what the *incremental* validity is of subtest score interpretations when the variance of the total score is already taken into account. Clinical practitioners may ask whether improvement in clinical decisions is achieved with the additional information from the group factors. There are a number of methods and study designs that can be used to determine the increase in validity, diagnostic efficiency, and treatment utility of group factor interpretations (Hunsley & Meyer, 2003; for examples of such study designs, see Nelson-Gray, 2003).

In recent years, the bifactor model has become more popular in clinical research, and several new computer programs have been developed to analyze data using the bifactor model (e.g., within an item response theory framework; Cai, Thissen, & du Toit, 2011; Cai et al., 2011b). Although we think that bifactor modeling is an interesting psychometric tool to investigate the data structure of psychological measures, there are also some issues that need further attention. First, in the present study we found two sources of variance due to clustered item content after the general factor was accounted for (a cluster of Items 3, 5, 7, 8, and 14 and a

cluster of Items 15 and 20). Although we concluded that these clusters did not support the creation of subscales, future research is needed to propose benchmarks for the value of ω_h or subscale scores and the percentages of common variance that are required for a group factors in a bifactor model to be considered meaningful (see also Cook et al., 2009; Sinharay, Puhan, & Haberman, 2010).

Second, not enough is known about the degree to which the cross-loadings of items bias the bifactor model parameters. Reise et al. (2010) discussed that cross-loadings can positively bias the factor loadings on the general factor and negatively bias the group factor loadings in a bifactor model (see also Asparouhov & Muthén, 2009). To identify the cross-loadings for the bifactor models, we conducted an exploratory bifactor analysis for a two-and three-factor solution using the Schmid–Leiman procedure (for an explanation of this procedure. see Reise et al., 2010; Schmid & Leiman, 1957). We used a polychoric correlation matrix with the Schmid routine included in the psych package (Revelle, 2012) of the R software program (R Development Core Team, 2008). We found no crossloadings with factor loadings larger than .30, but with a lower bound of .20 we found crossloadings for two items (1 and 13) in the two-factor solution and three items (2, 12, and 13) in the three-factor solution. To investigate the magnitude of the bias of these crossloadings for the bifactor Models D and E (for Model F, Items 1 and 13 were not part of the group factors), we added the item cross-loadings to the models. For the adjusted Model D, factor loadings for Item 1 were G = .86, g_{CA} = -11, g_{SV} = .00; and for Item 13, G = .56, g_{CA} = .23, g_{SV} = .22. For the adjusted Model E, factor loadings for Item 2 were G = .67, g_C = .26, g_A = -.02; for Item 12, G = .72, g_S = .20, g_A = .04; and for Item 13, G = .57, g_C = .23, g_S = .18. Note that the factor loadings on the group factors were very low. For most items the cross-loadings positively biased the factor loadings on the general factor, but not much. The impact of these changes on the explained variance of the general factor was small; for the adjusted Models D and E, the ECV of the general factor was still larger than 71%, and the interpretations of our results did not change.

In sum, with the current results in mind, we recommend that researchers use an investigative approach, such as a bifactor analysis, when evaluating the factor structure of psychological measures (with data that demonstrates construct-relevant multidimensionality) to inspect the extent to which the variance in item responses is due to a general or group factor. Furthermore, the take-home message for clinical practitioners is that they can readily

interpret the BDI–II total score as an estimate of the overall severity of depression, but they should be careful when interpreting subscale scores as if these scores were unique and reliable.

Chapter 5
Measuring Individual Significant Change on the BDI-II through IRT-based Statistics

Abstract

Several studies have emphasized that item response theory (IRT)-based methods should be preferred over classical approaches in measuring change for individual patients. In the present study we discuss the advantage of a simple IRT-based statistical test (Z-test) to study whether individual change on the BDI-II is statistically significant and we compare results obtained with the Z-test to those obtained by the Reliable Change Index (RCI) in a sample of clinical outpatients. Mean group differences between the Z-test and the RCI were similar, but results differed for individuals with pretest-posttest score differences that were close to statistical significant change. We show that this may have important consequences for the measurement of change in psychotherapy outcome research and clinical practice.

This chapter has been resubmitted as:

Brouwer, D., Meijer, R. R., & Zevalkink, J. Measuring Individual Significant Change on the BDI-II through IRT-based Statistics. *Psychotherapy research.*

5.1 Introduction

The general aim of every psychotherapeutic treatment is to reduce a patient's emotional distress, change the beliefs about the self, and, sometimes, try to increase moral behavior. The evaluation of change during psychotherapeutic treatments concerns patients, therapists, and policymakers alike and is of critical importance to assess the effect of a clinical treatment and to monitor the trajectory of a patient. For example, Lambert et al. (2003) showed in a meta-analysis that monitoring patient treatment responses and providing feedback to the therapist about the patient's progress significantly improved psychotherapy outcome compared to providing no information about the patient's progress during treatment. Therapy outcome results improved when therapist were given specific feedback about patients who were expected to leave treatment before receiving therapeutic benefit or who were thought to be at risk of having a negative treatment outcome. Lambert et al. (2003) stated that this type of independent patient information may prompt psychotherapists to modify their treatment approach and, for instance, change the treatment intensity or refer the patient to other health care providers (Lambert, 2007; Percevic et al., 2006).

To measure individual change and to help practitioners to determine whether significant change has occurred high quality measurement instruments and sound statistical methods are needed. Numerous studies have been devoted to the measurement of individual change and it has been a topic of furious debates (e. g., Cronbach & Furby, 1970; Collins, 1996; Williams & Zimmerman, 1996, Mellenbergh & van den Brink, 1998). Some of the problems with the traditional approaches to measure change are due to the shortcomings of the underlying statistical theory of these approaches, which is classical test theory (CTT). In recent years, the use of item response theory (IRT, Embretson & Reise, 2000) has started to replace CTT in test construction and test evaluation. In a number of psychometric oriented studies, authors have called for the use of IRT-based methods to study individual clinical significant change instead of traditional methods based on total scores (e.g., Doucette & Wolf, 2009; Finkelman, Weiss, & Kim-Kang, 2012; Reise & Haviland, 2005; Reise & Henson, 2003; Santor & Ramsay, 1998; Thomas, 2011b). In a special issue of *Psychotherapy Research* on quantitative and qualitative methods Doucette and Wolf (2009) explained in detail the advantages of IRT over CTT for psychotherapy research. In the field of psychotherapy and clinical psychology

research, however, there are almost no studies that use IRT-based statistics to assess clinical change.

The aim of the present study is to fill this gap and to compare a routinely used traditional approach with an IRT-based test statistic to establish whether a patient's pre- and posttest scores show statistically significant change (improvement or decline). Doing this, we follow the advice of Doucette and Wolf (2009) who questioned the measurement precision of many clinical instruments used in psychotherapy research. More specifically, the aim of this paper is (1) to use IRT methods to investigate how particular scale characteristics may influence the way clinicians demonstrate therapy effects and (2) to compare differences in results obtained from the Reliable Change Index (RCI; Jacobson & Truax, 1991) with those from a Z-test (Guo & Drasgow, 2010) based on IRT. To do this, we used data from a sample of clinical outpatients that completed the often-used BDI-II (Beck et al., 1996) before and during or after treatment.

5.1.1 Measurement of individual significant change: clinical and statistical significant change

There is a distinction between clinical and statistical significant change. First, the concept of *clinical* significant change can be approached from different angles. From a methodological perspective, change can be considered clinically significant when the measurement after treatment falls within the normative range on relevant outcome measures (Jacobsen, Follette, & Revenstorf, 1984; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999). There are numerous ways to assess clinical significance methodologically (e.g., Atkins, Bedics, McGlinchey, & Beauchaine, 2005; Bauer, Lambert, & Nielsen, 2004; Lambert & Ogles, 2009; Jacobsen, Roberts, Berns, & McGlinchey, 1999; Ogles, Lunnen, & Bonesteel, 2001; Wise, 2004). From a therapist and patient perspective, clinically significant changes refer to changes in psychological functioning that are meaningful for the patient and improve their well-being (e.g., feeling less depressed, experiencing changes in behavioral patterns contributing to suffering, or feeling more secure in intimate relationships; Binder, Holgerson, & Nielsen, 2010; Valkonen, Hänninen & Lindfors, 2011). Whether the patient and therapist consider the change to be clinically significant or meaningful influences the decisions they make about the treatment. For example, they might want to discuss a change in frequency or type of therapy when behavior patterns do not change during the course of treatment.

Second, although clinicians act primarily upon clinical considerations they sometimes also take the scores on clinical scales into account. For example, a therapist might want to terminate treatment when a patient's scores on relevant questionnaires are much lower as compared to pre-treatment scores and fall in normative ranges. In that case, they depend on the statistical methods that are used to estimate whether the change is not only clinical but also statistically significant. For change on clinical scales to be considered clinical significant, it should be at least statistically significant so that changes in observed scores reflect real changes rather than measurement error. To this end, Jacobson et al. (1984; Jacobson & Truax, 1991) defined a change index, the RCI, that expresses whether changes in observed scores reflect real changes rather than measurement error.

5.1.2 The RCI-index

The RCI is often used to define whether the patient's change, through therapeutic intervention, is statistically significant. Wise (2004) discussed that several variations of the RCI have been proposed, but that after comparing these methods multiple authors recommended the Jacobson and Truax (1991) method (e.g., Bauer et al., 2004; Maassen, 2001; McGlinchey, Atkins & Jacobson, 2002; Speer & Greenbaum, 1995). Furthermore, Ogles, Lunnen, and Bonesteel (2001) found that, in a selected review of the RCI literature including 74 published studies the Jacobson and Truax method (1991) was most often used. This RCI (Christensen and Mendoza, 1986; Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991) is defined as the ratio of an individual's observed change and the standard error of measurement of the difference score,

$$RCI = \frac{X_2 - X_1}{\sqrt{2 \times \left(SD_1 \times \sqrt{(1 - r_{xx'})}\right)^2}},$$

where X_1 = pretest score, X_2 = posttest score, SD_1 = standard deviation of scores in the pretest sample and $r_{xx'}$ = reliability. Clinicians use the RCI to calculate how much change in observed score units represents statistically significant change for a particular patient. In large samples the difference between scores follows a standard normal distribution, depending on the significance level that is required different cut-off points for the RCI can be used (see Jacobson & Traux, 1991). For example, an RCI of 1.96 reflects a significance level at a Type

I error rate $\alpha = .05$ for rejecting the null hypothesis that there is no change between measurements.

The standard error of the difference score (the denominator in the RCI formula) is often defined as a function of the standard deviation of scores in the pretest sample and the test-retest reliability (Jackobson & Traux, 1991). A basic assumption of the RCI is that the measurement precision across person's scale scores is the same. However, this is often not the case. Studies using IRT showed that for many clinical instruments, the measurement precision is lower at the scale levels that indicate the absence of distress or non-problematic behavior, and higher at scale levels that indicate distressed or problematic behavior. This is due to the fact that many clinical scales are unipolar (e.g., Reise & Waller, 2009; Meijer & Egberink, 2011). That is, these scales consist of items that indicate a certain amount of distress (or other clinical relevant constructs), whereas items that indicate more healthy symptoms are left out. As a result, these unipolar scales are not very sensitive to differences in scores at one end of the scale continuum. IRT-based methods do take into account the difference in measurement precision across the scale as we will describe next, after we first discussed some of the relevant basics of IRT.

5.1.3 Item Response Theory

IRT models are based on the idea that psychological constructs such as depression are latent, that is, they are not directly observable. Knowledge about latent constructs can be obtained through the observed responses of persons to a set of items that measure some aspect of that construct (e.g., Doucette & Wolf, 2009; Embretson & Reise, 2000; Meijer & Baneke, 2004). In IRT, the relation between a person's latent variable estimate (theta or θ; severity of depression in this study) and the probability that a person gives a particular item response can be described through a specific model. In practice, this θ value is always estimated. The θ scale is in a standard score metric with mean score of 0 and a standard deviation of 1. Thus, θ = 1 implies that a patient scored one standard deviation above the mean score in the standard score metric. An often-used IRT model for polytomous item scores is the graded response model (GRM, Samejima, 1969; 1997). The GRM is suitable for analyzing ordered response categories, such as likert-type rating scales and several researchers used this model to analyze clinical scales (e.g., Cole et al., 2011; Emons et al., 2007; Purpura, Wilson, & Lonigan, 2010; Walters et al., 2011; Wu et al., 2012).

In the GRM items are defined by a discrimination parameter (*a*) and two or more location parameters (*b*). The magnitude of the discrimination parameter reflects the degree to which the item is related to the θ. This means that for high *a*-values the response categories accurately differentiate among θ levels. The location parameters reflect the spacing of the ordered response categories along the θ-scale. The location parameter *b* for category *m* can be interpreted as the point at the latent scale where there is a 50% chance of scoring in category *m* or higher (in this study *m* = 1, 2, or 3). The BDI-II items have four response categories, but the location parameters are only given for response categories 1, 2, and 3 because the probability of choosing category 0 or higher equals unity. Thus, respondents with a θ-value higher than b_2 have more than 50% chance of responding in category 2 or 3. These probabilities can be used to determine the category response functions (CRF), which describe the probability of responding in a particular response category conditional on θ.

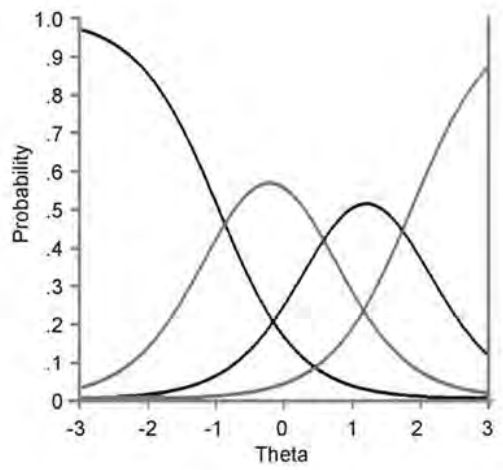

Figure 5.1: The Category Response Functions for Item 20 of the BDI-II. The Category Response Functions (CRF) for Item 20 (Tiredness or fatigue) were estimated under the GRM. The points where the CRF's cross are, from left to right, the item location parameters b_m (for *m* = 1, 2 or 3) where there is a 50% change of responding in category *m* or higher.

As an example, Figure 5.1 shows the CRFs for Item 20 (Tiredness or Fatigue) of the BDI-II. As can be seen, the probability of responding to a particular response category is conditional on θ. The item discrimination parameter determines the steepness of the curves, the item location parameters determine where the curves intersect. For Item 20, the first intersection at b_1 = -.98 indicates that for θ = -.98 the chance is 50% of responding with either 0 'I am no more tired or fatigued than usual' or 1 'I get more tired or fatigued more easily than usual'. For higher θ values it is more likely that a person responds with a value larger than 0.

When comparing the RCI, based on classical test theory, with an IRT-based method for measuring statistical significant change, two distinctions between the two theories are most relevant. A first important distinction is that in CTT usually the sum of item scores is used to scale persons. This method is almost always used in clinical practice, mainly because of its simplicity. In IRT, however, item scores are weighted by the item characteristics (e.g., persons who score higher on discriminating items receive higher estimates). Consequently, persons with the same unweighted total score may obtain different θ values depending on their item score pattern. IRT-based weighted item sums provide more information than unweighted item sums. In fact, Dumenci and Achenbach (2008) demonstrated that unweighted item sums "may seem like a simple solution, but it invites measurement inaccuracies, especially in both tails of the distributions" (p. 61; see also, Thomas, 2011b).

A second important distinction between CTT and IRT is that in the former, total scores on a test are assumed to have equal standard errors regardless of their numerical values, whereas in IRT the standard errors for different values of θ may differ. In CTT, there is one reliability estimate. In parametric IRT, the concept of reliability is replaced by the concepts of item and test information (Embretson & Reise, 2000). Information is a psychometric concept that indicates how well an item differentiates among persons who are at different levels of θ (information is estimated for every θ estimate). In general, items with larger discrimination parameters provide relatively more information and the item location parameter determines where that information is located. Item information is additive across the items administered and test information is maximized around the location parameters. The standard error of a θ is inversely related to the test information function. This means that when test information is large, the standard error is small. In short, θ estimates may have different standard errors depending on how discriminating a set of items is in different ranges of θ.

5.1.4 Reliability versus Information

A fundamental issue that hardly if ever is discussed in the psychotherapy and clinical psychology literature and that is a common misconception is that low reliability implies low measurement precision and hence imprecise statements on individual changes. Several authors noted that low reliability does not imply a lack of precision per se (e.g., Collins, 1996; Williams & Zimmerman, 1996; Mellenbergh & van den Brink, 1998). To understand this, it is important to further distinguish between two aspects of measurement precision, namely reliability and information. Reliability is a population dependent concept of measurement precision, as it directly depends on the variability of the θ estimates in the population. In contrast, information only depends on the estimate of θ for the individual of interest, and hence is person dependent, not population dependent. This means that a measurement instrument that cannot detect inter individual differences with a satisfactorily precision, may do so with respect to intra-individual change (Mellenbergh, 1999). If reliability is low, but for a given person information is high then statements on population change are imprecise, but statements on the person's change are precise.

When we apply this knowledge to change measures such as the RCI, reliability expresses the measurement precision in studying *population* change. IRT provides us with appropriate tools for the measurement precision at the individual level. Therefore, it seems better suited to investigate clinical change at the individual level. Information can be calculated easily in the context of IRT models. Knowledge about which scores ranges for clinical scales (in this study the BDI-II) have high or low measurement precision can be used to improve the measurement of change. It is also important to note that as individuals improve as a result of treatment, they move toward the lower end of the BDI-II scale indicating milder levels of distress, where the measurement model is often less precise, information is lower here and thus the standard error for these scores is larger.

5.1.5 The Z-test

The Z-test is an IRT based change index. With the Z-test the numerical θ -values of persons at different time points can be compared. Because the Z-test is based on IRT the metric that is used is not based on raw scores, but on θ estimates. Suppose that we want to compare θ_1 (pretest score on the latent variable scale) with θ_2 (posttest score on the latent variable scale).

One possibility is to test the null hypothesis H_0: $\theta_1 = \theta_2$ (no change) against H_a: $\theta_1 \neq \theta_2$. Guo and Drasgow (2010) provided the following test statistic:

$$Z = \frac{\theta_2 - \theta_1}{\sqrt{\mathrm{SE}_2^2 + \mathrm{SE}_1^2}},$$

where θ_i is the maximum likelihood estimate (MLE) of θ_i and SE_i is the standard error of estimation for a given θ_i (for i =1,2). When the abilities are estimated using maximum likelihood or the Bayesian method, the estimates are asymptotically normal (Bock & Mislevy, 1982). Thus, θ_1 and θ_2 are approximately normal given sufficient test length, and the score difference as θ_2 - θ_1 should also follow a normal distribution. With the IRT property of conditional independence under H_0, θ_1 and θ_2 will be independent. So under H_0, the standardized score difference between the two tests follows a standard normal distribution. H_0 is rejected for a fixed Type I error rate α if $|Z| \geq z_{1-\alpha/2}$. Like the RCI, a Z-score of 1.96, reflects with relative certainty ($\alpha < .05$) that actual change has occurred. There are several methods for detecting change within an IRT framework. Guo and Dragow (2010) compared the Likelihood-Ratio-test (LR-test) with the Z-test for their power to detect change in θ levels. The results of their simulation study demonstrated that the "*Z-test is simpler and more effective than the LR-test*" (p. 362). In the context of computerized adaptive testing, Finkelman et al. (2010) also compared different methods for assessing statistical significant change with IRT in a simulation study. They concluded that a slightly different version of the Z-test given in Equation (2) exhibited the highest power to detect change and the lowest type I error rate over different amounts of change conditions ($0.5 < \Delta\theta < 1.5$) as compared to a test based on overlapping confidence intervals and the LR-test.

In the present study, we investigated the particular scale characteristics that influence the measurement precision of scores on the BDI-II in a sample of clinical outpatients at two measurement points. We compared the number of patients that were considered statistically significantly changed using the RCI and the Z-test. In addition, we investigated specific cases in which change was statistically significant for one index, but not for the other.

5.2 Method

5.2.1 Measure and Participants

Measure. The Beck Depression Inventory II (BDI-II; Beck et al., 1996) is a 21-item self-report questionnaire that is used to assess the severity of depressive symptoms that correspond to the Diagnostic and Statistical Manual of Mental Disorders criteria for major depressive disorder (DSM-IV-TR; American Psychiatric Association, 2000). Each item consists of four statements, which are scored from 0 through 3. For example, Item 2 "Pessimism" has four response categories ranging from 0 "I am not discouraged about my future" through 3 "I feel my future is hopeless and will only get worse". The BDI-II is used to assess the severity of a patient's depression before clinical treatment for diagnostic purposes and with intervals during and after treatment to detect treatment progress or treatment stagnation. Its psychometric properties have been found to be satisfactory in several studies (e.g., Beck et al., 1996; Steer et al., 2002).

Participants. The sample consisted of 104 outpatients, 76% females and 24% males. Participants signed informed consent and completed the BDI-II on two occasions, during the intake procedure (pretest sample) of a clinical treatment and approximately a year later (posttest sample; in months: $M = 15.2$, $SD = 6.0$). The data were obtained as part of a routine outcome monitoring project for persons in treatment in the period between the years 2009 and 2012 at a community mental health clinic specialized in ambulatory psychoanalytic treatment. At intake, the mean age was 33.6 years (SD = 10.1) for the entire sample. 61.5% of the outpatients reported Dutch as their dominant culture when asked in which culture they were raised and 62.4% had a Bachelor or Master degree.

Psychiatric diagnoses were assessed at intake in a consensus meeting and patients were classified according to the DSM-IV-TR. For 99.0% of the 104 participants DSM-IV-TR Axis I classifications were available. Most frequent classifications of clinical syndromes were mood disorders (53.4%), anxiety disorders (24.3%), adjustment disorders (13.6%) and eating disorders (3.9%). In addition, on Axis I of the DSM-IV-TR additional problematic conditions were classified that could not be classified as clinical syndromes, but were serious enough to warrant independent clinical attention. Most frequent classifications of additional problematic

conditions were partner relational problems (21.4%), identity problems (31.1%), and phase of life problems (14.6%).

Most frequent classifications of personality disorders on Axis II were personality disorder not otherwise specified (50.5%), no personality disorder (17.5%), avoidant personality disorder (28.2%), dependent personality disorder (14.6%), diagnosis postponed (10.7%), narcissistic personality disorder (9.6%) and obsessive-compulsive personality disorder (8.7%). Note that multiple classifications on Axis I and II were possible and that a part of the Axis II classifications that were postponed at intake were diagnosed on a later occasion between the two measurements.

After intake, patients were assigned to different psychotherapeutic treatments: psychoanalytic psychotherapy (49.0%), short term psychoanalytic treatment (21.2%; McCullough, Kuhn, Andrews, Kaplan, Wolf, & Hurley, 2003), psychoanalysis (11.5%), mentalization based treatment (8.7%; MBT, Bateman & Fonagy, 2004), psychoanalytic group psychotherapy (7.7%), transference-focused treatment (1,0%; TFP, Clarkin, Yeomans, & Kernberg, 2006) and parental support treatment (1,0%). Between intake and second measurement 24 of the 104 patients (23.1%) finished treatment, 17 of these 24 patients finished a short-term psychoanalytic treatment.

5.2.2 Analyses

Descriptive Statistics and IRT Parameters. First, we investigated the traditional descriptive statistics for the BDI-II items in the current sample pre- and posttest. Second, we estimated the items discrimination and location parameters for the GRM using Multilog 7.0 (Thissen et al., 2003). For calibration of GRM item parameters we used a sample of N = 1530 clinical outpatients from the Brouwer, Meijer, and Zevalkink (2012) study. The current sample was drawn from the same population. Prior to conducting an IRT analysis, we determined whether the data were suited for IRT modeling. The GRM assumes that the data are unidimensional, implying local independence of item responses after controlling for a single common factor, and that the IRFs are monotonically increasing (e.g., Embretson & Reise, 2000; Reise & Haviland, 2005). With MSP 5.0 (Molenaar & Sijtsma, 2002) we found no significant violations of monotonicity. Furthermore, Brouwer et al. (2012; see also Al-Turkait & Ohaeri, 2010; Osman et al., 2008; Quilty et al., 2010) and Ward (2006) demonstrated that

the BDI-II items showed some local dependence, but that there was a large common factor that explained most of the common variance in BDI-II scores. Brouwer et al. (2012) concluded that the BDI-II data can be considered unidimensional for practical purposes, that is, without a significant distortion of item (and factor) parameters due to multidimensionality. Third, we transformed the test information curves (also provided by Multilog 7.0) into a standard error curve. Scale information is inversely related to the conditional standard error of measurement, that is:

$$SE|\theta = \frac{1}{\sqrt{I|\theta}}.$$

The standard error curve gives the standard error for different scores of θ. Fourth, we created the scale response curve (SRC; e.g., Reise & Haviland, 2005) with the calibrated item parameters. The SRC describes the relation between θ scores and the expected (weighted) raw scores on the BDI-II. As described above, the CRF's predict the probability of a person responding in a particular category, based on that person's θ estimate (P_{im} (θ); see pp. 98-99 of Embretson & Reise, 2000). For every θ estimate we calculated the expected response for each item $E(x_i) = P_{i1}(\theta) + 2P_{i2}(\theta) + 3P_{i3}(\theta)$ Reise and Haviland (2005) demonstrated that inspection of the SRC is informative for the study of change. For example, for a cognitive problem scale they found that θ had a nonlinear relationship to the raw scale scores and they concluded that equal raw score differences had different implications in terms of the change on θ measured by the instrument. Fifth, using Multilog 7.0 we calculated the θ estimates and its standard error for all persons in the current sample (based on the calibrated item parameters). So, in addition to the SRC, we also inspected scatter plots of the actual correlation between θ estimates and unweighted total scores for the current sample at pre- and posttest administrations.

RCI and Z-test. Sixth, we calculated RCI and Z-test scores for all persons. For both change indices we compared the number of patients for whom change was considered statistically significant. For both indices we used 1.96 as a cut-off value: scores smaller than -1.96 or larger than +1.96 were considered to reflect statistical change. The RCI and Z-test values can be used to categorize the change of scores for a person. Jacobson and Truax (1991) proposed the following classification of clinical change scores: recovered (passed both clinical significant and statistically significant criteria), improved (only passed statistically significant

criterion), unchanged (passed neither) or deteriorated (passed statistically significant criterion in wrong direction). Because we investigated statistically significant change and not clinical significant change, in this study only the latter three classifications were used.

The main differences between the RCI and the Z-test are (1) the metric on which they are based (raw scores versus θ estimates) and (2) the method for calculating the standard error of the difference score (based on reliability versus information). For the Z-test the standard error is based on the standard errors of the estimated θ scores on both pre- and posttest. For the RCI the standard error is defined as a function of the standard deviation in the population and the test reliability. We used Cronbach's α = .90 as an estimate for the reliability[1]. Finally, we investigated specific cases in which change was considered statistically significant for one index, but not for the other.

[1] Hiller, Schindler, and Lambert (2012) pointed out that the authors of the RCI did not exactly specify which reliability value researchers should use in the RCI formula and that internal consistencies and not retest values are preferred, but that internal consistency values in many cases differ over studies. The choice of reliability coefficient therefore seems rather arbitrarily. The BDI-II total score reliability of at least α = .90 is often reported in the research literature within clinical samples (e.g. Arnau et al., 2001; Beck et al., 1996; Beck et al., 2002; Buckley et al., 2001; Dozois et al., 1998; Osman, Downs, Barrios, Kopper, Guttierrez, & Chiros, 1997; Steer et al., 1999) and corresponds with Cronbach's alpha values found in the current sample. A more conservative value for the reliability of summed total BDI-II scores, based on the bifactor results, ω_h = .85 was recommended by Brouwer et al. (2012). The comparison between RCI and Z-test would not be fair if we used the reliability estimate of the RCI based on this more stringent reliability index and not for the Z-test. Bifactor IRT analysis methods for polytomous item responses are still in an experimental stage.

5.3 Results

5.3.1. Descriptive statistics

Table 5.1 presents descriptive statistics for the item and scale scores of the pretest and posttest scores of the BDI-II. Cronbach's α equaled α = .90 (95% CI = .87 - .93) for pretest and α = .91 (95% CI = .88 - .93) for posttest summed BDI-II total scale scores.

Table 5.1

Descriptive Statistics for the BDI-II Items for Pretest and Posttest in a Sample of N = 104 Clinical Outpatients.

		Pretest					Posttest				
Nr	Content	*M*	*SD*	r_{it}	Skewness	Kurtosis	*M*	*SD*	r_{it}	Skewness	Kurtosis
1	Sadness	.66	.53	.49	-.10	-.88	.46	.57	.66	.76	-.46
2	Pessimism	.94	.85	.52	.78	.15	.63	.74	.57	.85	-.20
3	Past failure	.72	.82	.47	.65	-.91	.58	.84	.56	1.10	-.15
4	Loss of pleasure	.83	.82	.75	.43	-1.10	.50	.65	.60	1.14	.98
5	Guilty feelings	.88	.81	.56	.44	-.78	.67	.81	.57	.98	.18
6	Punishment feelings	.51	.97	.36	1.67	1.33	.38	.79	.49	2.05	3.13
7	Self-dislike	1.13	.86	.60	.13	-.96	.64	.81	.60	.83	-.66
8	Self-criticalness	1.12	.99	.57	.25	-1.20	.71	.87	.53	.94	-.14
9	Suicidal thoughts	.33	.51	.28	1.15	.20	.26	.48	.34	1.58	1.54
10	Crying	.92	1.00	.44	.90	-.28	.46	.74	.42	1.79	3.16
11	Agitation	.76	.66	.53	.50	.09	.48	.64	.56	1.18	1.22
12	Loss of interest	.67	.81	.65	1.20	1.09	.49	.65	.66	1.17	1.05
13	Indecisiveness	.85	1.02	.60	.90	-.46	.63	.95	.69	1.41	.84
14	Worthlessness	.67	.81	.67	.65	-1.18	.55	.79	.63	.97	-.71
15	Loss of energy	1.00	.86	.65	.27	-1.02	.66	.77	.63	.77	-.51
16	Changes in sleep	.89	.91	.41	.81	-.17	.78	.85	.40	1.09	.78
17	Irritability	.78	.75	.56	.65	-.11	.51	.65	.60	.90	-.34
18	Changes in appetite	.62	.80	.40	1.35	1.43	.43	.62	.31	1.59	3.46
19	Concentration difficulty	.93	.87	.69	.30	-1.24	.61	.77	.60	.79	-.89
20	Tiredness or fatigue	.90	.84	.56	.47	-.76	.63	.75	.52	1.00	.35
21	Loss of interest in sex	.50	.79	.35	1.35	.72	.32	.60	.45	1.95	3.84
	Total	16.61	10.19		α = .90; ω_t = .92		11.37	9.23		α = .91; ω_t = .93	

Note: r_{it} = corrected item-test correlation. In the last row the total scale *M, SD*, Cronbach's α and McDonald's $omega_{total}$ are provided

The corrected item-total correlations ranged from r_{it} = .30 through r_{it} = .67. Overall, the item mean scores were lower as compared to BDI-II scores reported in other clinical samples but higher as compared to normal samples, for example, those reported by Beck et al. (1996, 2002). The item mean scores for the posttest scores were lower than those of the pretest scores; the decrease of item means for Item 7 (Self-dislike), Item 8 (Self-criticalness), and Item 10 (Crying) were largest.

Table 5.2

IRT Parameters for the BDI-II Items in the N = 1530 Clinical Outpatients Calibration Sample

		IRT parameters			
Nr	Content	*a*	b_1	b_2	b_3
1	Sadness	2.17	-.76	1.41	2.16
2	Pessimism	1.71	-.87	.93	1.94
3	Past failure	1.37	-.39	.63	2.64
4	Loss of pleasure	1.85	-.76	.76	2.21
5	Guilty feelings	1.21	-.73	1.37	2.92
6	Punishment feelings	.97	.74	1.79	2.48
7	Self-dislike	1.55	-1.09	.33	2.02
8	Self-criticalness	1.28	-1.07	.30	2.34
9	Suicidal thoughts	1.40	.22	2.91	4.40
10	Crying	.99	-.87	1.14	1.85
11	Agitation	1.25	-1.04	1.62	2.79
12	Loss of interest	2.02	-.51	1.07	1.84
13	Indecisiveness	1.47	-.60	.55	1.22
14	Worthlessness	1.78	-.29	.55	2.28
15	Loss of energy	1.85	-1.19	.19	2.15
16	Changes in sleep	1.04	-1.23	.70	2.05
17	Irritability	1.28	-.74	1.15	2.87
18	Changes in appetite	1.04	-.22	1.66	2.91
19	Concentration difficulty	1.74	-1.00	.21	2.19
20	Tiredness or fatigue	1.69	-.98	.54	1.88
21	Loss of interest in sex	.89	.26	1.79	3.59

Note: a = discrimination parameter (for parametric IRT scaling), b_m = location parameter; the point at the latent variable continuum where there is a 50% chance of scoring in category m.

The skewness and kurtosis for the posttest were higher as compared to the pretest indicating a larger tail to the right and a higher peak, due to more scores in the lower ranges of the scale.

5.3.2 Standard Error

Table 5.2 shows the location and discrimination parameters for the BDI-II items estimated under the GRM in the calibration sample. Item 1 (Sadness), Item 12 (Loss of interest), and Item 15 (Loss of energy) had the highest discrimination parameters (ranging from a = 2.17 through a=1.85, respectively) and Item 21 (Loss of interest in sex) had the lowest discrimination parameter (a = .89). The item location parameters ranged from -1.23 through 4.40, with the highest value (b_3 = 4.40) for Item 9 (Suicidal thoughts or wishes). This high value reflects the extreme statement 'I would kill myself if I had the change' which is indicative of severe depressive symptomatology. All other item location parameters ranged from 1.23 through 2.92. In this range most information is located and, consequently, in this range the standard error is relatively small.

Figure 5.2 shows the standard error across different scores. The standard error curve demonstrated that for θ score estimates ranging from θ = -1 through θ = 3 the measurement precision was highest.

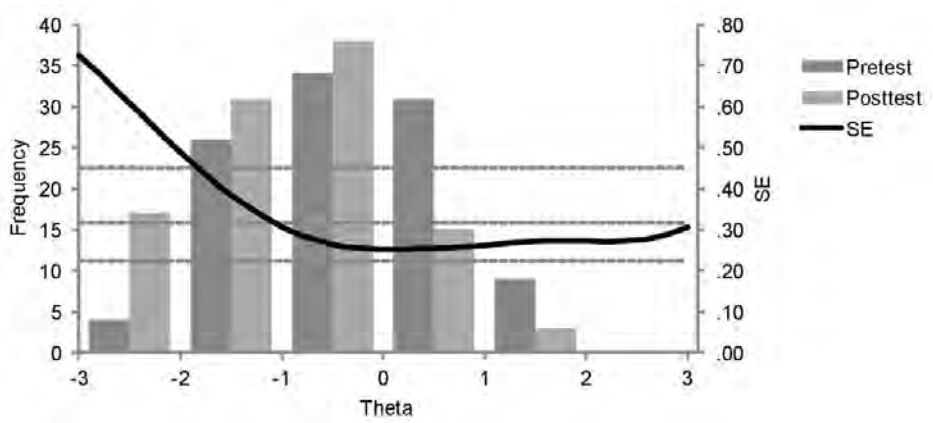

Figure 5.2: The Standard Error curve and frequency of score estimates for two measurements of the BDI-II in a sample of N = 104 clinical outpatients. The bars represent the frequency of score estimates; The lower, middle and upper dashed horizontal lines showing the amount of standard error corresponding to a reliability of .95, .90, and .80 respectively.

The standard errors in this range corresponded with a reliability coefficient between .90 (middle horizontal line) and .95 (lower line). The standard errors were larger for lower θ score estimates and consequently the reliabilities for these scores were lower, crossing the upper horizontal line that represents a reliability of .80. The items covered a large range of the θ scale with high measurement precision. However, Figure 5.2 also shows that many θ score estimates for the pre- and posttest samples fell outside this region. From the frequency distribution it can be deduced that 30 score estimates (29%) for the pretest sample, and 48 score estimates (46%) for the posttest sample were located in the region of the θ continuum where the least information was located. Consequently, these scores had a standard error larger than .32, which corresponded with reliability coefficient smaller than .90.

5.3.3 Relation between latent variable and raw score metric

Figure 5.3 shows the SRC and two scatter plots of the unweighted total scores and θ estimates and illustrates two important results. First, the SRC demonstrated that the relation between θ estimates and expected total scale scores was nonlinear. This means that an equal change on θ did not produce an equal change on the entire raw score metric, indicating that the raw score metric is not suited for interval level measurement. As can be seen, an expected total score change of 10 from 35 to 25 (from severe to moderate depression) corresponded with a change of $\Delta\theta = .82$ (from $\theta = 1.20$ to $\theta = -.38$), but an expected total score change of 10 from 15 to 5 (from mild to minimal depression) corresponded with a change $\Delta\theta = 1.19$ (from $\theta = -.50$ to $\theta = -1.69$). Second, the two scatter plots demonstrated that one θ estimate corresponded with multiple unweighted total scores. These results indicated that at least two scale characteristics were relevant with regard to the measurement precision of scores and the measurement of statistical significant change with the BDI-II. First, we found that for the mild, moderate, and severe depression categories the measurement precision was high, but for the minimal depression category (more specifically for θ estimates below -1) measurement precision was low.

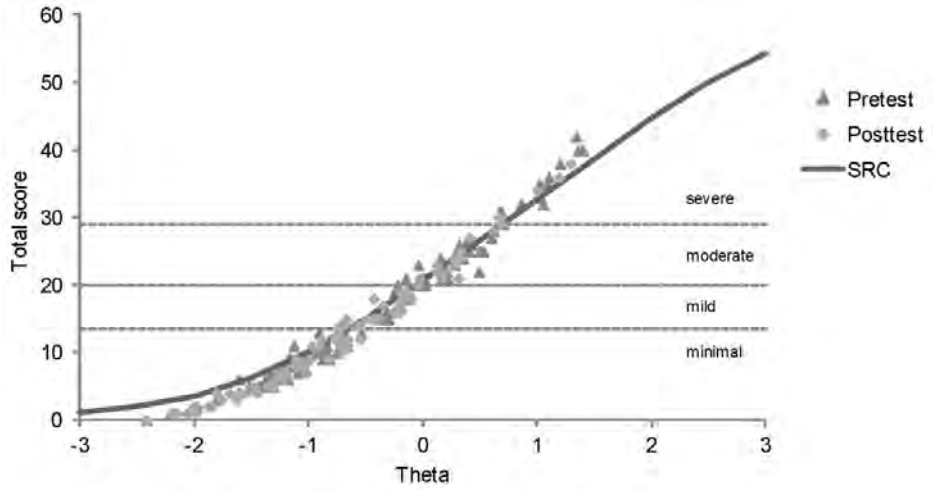

Figure 5.3: The Scale Response Curve combined with two scatter plot of the relation between theta and raw score for two measurements of the BDI-II in a sample of N = 104 clinical outpatients. SCR = Scale Response Curve. The SRC is a function of theta and *expected* (weighted) total scores. The scatter plots of pre- and posttest scores represent the correlation between theta and *observed* unweighted total scores. The dotted lines separate the four depression categories that were defined by Beck et al. (1996); 0–13: minimal depression; 14–19: mild depression; 20–28: moderate depression; and 29–63: severe depression.

Second, we found that (a) the unweighted total score was a less informative indicator of depression severity, because the same unweighted total scores indicated different latent variable levels and that (b) the raw score metric was not suited for interval measurement, because the same amount of raw score change related to different amounts of change on the scale. Next, we compared the change outcomes for the current sample as indicated by the RCI versus the Z-test.

5.3.4 Change Indices

On a group level, the absolute Z-test values (M = 1.89, SD = 1.43) were similar to the RCI values (M = 1.86, SD = 1.54, Cohen's d = .02). In 52% of the cases the absolute Z-test value was larger than the absolute RCI value (and in 44% of the cases |RCI| > |Z-test|). On an individual level we found that using the RCI, 33 patients improved, 65 did not significantly change, and six deteriorated. With the Z-test, 35 patients improved, 61 did not significantly change, and eight deteriorated. Using the RCI and the Z-test 96 persons were classified to the same change categories, but for eight persons results differed.

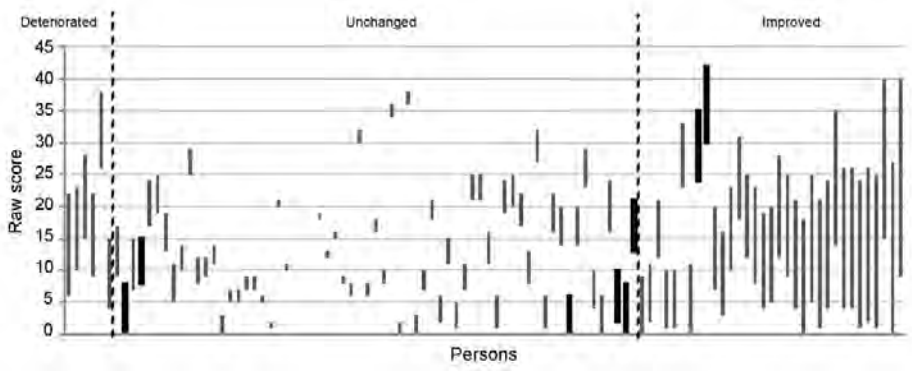

Figure 5.4: The raw difference scores for the BDI-II for all N = 104 persons. The length of the bars represent the change between pre and posttest in raw total score, the changes are ordered by their magnitude of raw difference score, ranging from $\Delta X = 16$ (deteriorated) on the left through $\Delta X = -31$ (improved) on the right side. The dashed vertical lines represent $|RCI| = 1.96$. The left vertical dashed line separates the deteriorated cases (on the left) from the unchanged cases, the right vertical dashed line separates the unchanged cases from the improved cases (on the right), according to the RCI. The fat black bars indicate change for six persons that the RCI flagged as 'unchanged', but the Z-test as 'improved' or 'deteriorated' and two cases (upper right) where the Z-test flagged the change as 'unchanged' while the RCI flagged the change as 'improved'.

Figure 5.4 shows that the Z-test assigned six persons to 'improved' or 'deteriorated', whereas using the RCI these persons were classified as unchanged and two persons were classified as 'improved' that did not improve according to the Z-test. With the RCI (with $\alpha = .90$) a raw score difference of at least $\Delta X = 9$ leads to an $RCI > 1.96$ and thus a change which was flagged as statistically significant. However, according to the Z-test a smaller raw score difference (ranging from $\Delta X = 6$ through 8) also lead to a statistically significant change for six persons.

For most persons in this sample there was no difference between the RCI and Z-test when identifying statistically significant change. However, there were some interesting cases when the raw score change was between 6 and 8 and the RCI and Z-test resulted in different classifications.

Table 5.3

Change Statistics for two Persons.

Person	X_1	X_2	θ_1	θ_2	$SE(\theta_1)$	$SE(\theta_2)$	ΔX	$\Delta\theta$	RCI		Z-test	
1	24	16	.15	-.24	.26	.26	-8	-.38	-1.76	Unchanged	-1.04	Unchanged
2	10	2	-.76	-1.88	.28	.38	-8	-1.11	-1.76	Unchanged	-2.36	Improved

Note. The comparison between person 1 and 2 demonstrated that for the same $\Delta X = 8$, $\Delta\theta$ had different values. The different Z-test values resulted in different classification of change between the RCI and Z-test.

Consider the scores of person 1 compared to the scores of persons 2 in Table 5.3. The change of raw scores and consequently the RCI scores were similar. For both persons 1 and 2 $\Delta X = 8$ and RCI = 1.76, indicating no significant change. However, the change in θ was larger for person 2 ($\Delta\theta$ = -1.32 and Z = 2.36) as compared to person 1 ($\Delta\theta$ = -.38 and Z = 1.04). Figure 5.5 shows that a decrease or increase of 8 points on the raw score metric corresponded to a larger change in θ for the low range of the scale for person 2 compared to a change at the middle region of the BDI-scale for person 1. For person 2 the difference between the two methods was relevant, with the RCI this person was classified as 'unchanged' and with the Z-test this person was classified as 'improved'.

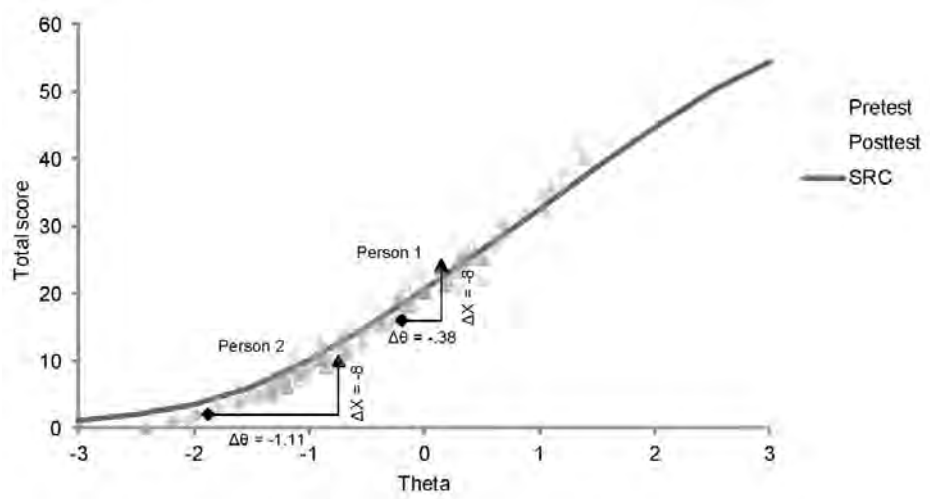

Figure 5.5: The pre and posttest raw total and latent variable scores for person 1 and 2. SCR = Scale Response Curve. The SRC is a function of theta and *expected* (weighted) total scores. The scatter plots of pre- and posttest scores represent the correlation between theta and *observed* unweighted total scores. The highlighted cases illustrate that equal raw difference scores can represent different latent variable difference scores

5.4 Discussion

Researchers have called for the use of IRT-based methods to study individual change in psychotherapy research instead of traditional methods based on total scores, but thus far there are almost no studies in this field that use IRT-based statistics to assess individual change. In the present study, we compared an IRT-based change statistic, the Z-test, with a traditional method, the RCI. We used pre- and post treatment scores of the BDI-II in a sample of clinical outpatients. In IRT measurement precision can differ across scores, that is, reliability is replaced by information that is conditional on the score estimates, and the metric for change is based on θ. Before we compared the two change indices, we therefore first investigated (1) the measurement precision conditional on θ and (2) the relation between the total score metric and the θ metric.

We observed that measurement precision was high (reliability between .90 and .95) for score estimates larger than θ = -1, which corresponded roughly with a raw score of 8 and larger. On the one hand this implies that the cut-off scores for different depression categories for the BDI-II (0–13: minimal depression; 14–19: mild depression; 20–28: moderate depression; and 29–63: severe depression), as defined by Beck et al. (1996) were located in a range of θ in which the BDI-II most precisely differentiated between scores. On the other hand, it was clear from our sample of clinical outpatients that a substantial proportion of scores fell outside this range. In fact, for the posttest scores almost half of the score estimates were lower than θ = -1 (or a raw score lower than 8). For these persons posttest depression scores were relatively inaccurate.

This observation has an important implication when using IRT-based difference scores statistics such as the Z-test. Many clinical scales are only quasi scales (e.g., Cole et al., 2011; Emons et al., 2007; Meijer & Egberink. 2011; Purpura et al., 2010; Reise & Waller, 2009; Walters et al., 2011), that is, scales that measure what Reise and Waller (2009) called a quasi-trait: "a unipolar construct in which one end of the scale represents severity and the other pole represents its absence (depression versus not depressed). This is in contrast to a bipolar construct, where both ends of the scale represent meaningful variation (depression versus happiness)". As discussed above, our BDI-II data were in line with this observation. As a result, when a therapy has had the effect that a person becomes less depressed or even have

no depression signs at all, his or her depression scores in terms of latent scores can only be measured in an unreliable way with the BDI-II. This conclusion is in line with the general message of Reise and Waller (2009) and Doucette and Wolf (2009, p. 385) that clinical scales often have "inadequate item coverage at the ability location for those persons". As some authors emphasized: It is quite difficult to write items that have sufficient spread of the location parameters, without "gliding away" from measuring the intended construct (here: depression). Thus, also for the BDI-II a general remark made by Doucette and Wolf (2009) applies, namely that the BDI-II is only useful to report difference scores and thus psychotherapy effects within a certain range of depression.

With respect to difference between the Z-scores and the RCI scores we conclude that for many cases these statistics lead to the same classification of therapy effects. This was mainly due to the fact that (1) total scores and θ estimates were linearly related in most of the score ranges, and (2) in most cases the difference between pretest and posttest scores was so large or so small that the change was obviously statistically significant or not significant. In 19 cases (18.3%) where the raw score change was ranging from 6 through 8 this was less obvious. For six of these cases there was a different classification for the RCI and the Z-test. It is for these borderline cases that it is most relevant to further develop and apply new psychometric approaches to determine whether change is statistically significant. Although, in general, IRT requires larger sample sizes and the software to estimate model parameters is less well-known among clinical practitioners, the use of θ estimates instead of unweighted total scores and the person-tailored reliability estimates based on information rather than on reliability, can improve the measurement of change. In a recent study Cole et al. (2011) demonstrated that "*using IRT-derived information about symptom severity and discriminability substantially enhanced precision in the measurement of depression severity*" (p. 827) as compared to using the unweighted total scores and reliability coefficients.

There were several limitations to the current study. First, change on a scale that measures a specific symptom, like the BDI-II, is largest when the target population is selected for large values on that symptom and psychological intervention is aimed at reducing these symptoms. In the current study the sample was heterogeneous with respect to depression scores (and mean scores were lower as compared to other clinical BDI-II studies), because our sample represented a population of persons that seek help in psychotherapy with different types of

symptoms (53.4% of the participants were classified with a mood disorder). Furthermore, the psychological interventions were not exclusively aimed at reducing depressive symptoms and 76.9% of the patients were still in therapy at time of the second measurement. In our view, this heterogeneous group of patients is representative for the group of patients that we encounter in daily psychotherapeutic practice. However, future studies with homogeneous samples of depressed patients, for whom pretest mean scores are higher, and with psychological interventions specifically aimed at reducing depressive symptoms may further investigate the differences between change indices. Second, Hiller et al. (2012) pointed out that the authors of the RCI did not exactly specify which reliability value researchers should use in the RCI formula and that internal consistencies and not retest values are preferred, but that internal consistency values in many cases differ over studies. The choice of reliability value therefore seemed rather arbitrarily. We chose a reliability of .90 for the RCI formula. In an earlier study Brouwer et al. (2012) demonstrated that 83%–86% of the variance of a summed total score of the BDI-II was attributable to the general factor, and that thus a reliability estimate of .85 for BDI-II total scores may be an alternative to a reliability estimate of .90. Because for the current study we compared two change indices based on a unidimensional measurement model, we did not want to 'punish' the RCI method with a lower reliability if we did not apply the same constraints to the Z-test. Third, the results in this study are based on the one-factor model for the BDI-II. Although this choice is supported by previous findings (Brouwer et al., 2012), we encourage future researchers to examine the advantages of multidimensional IRT applications such as bifactor model IRT analyses for the measurement of individual change (such methods are currently being developed and tested, e.g., Cai et al., 2011a;Cai et al., 2011b; Gibbons et al., 2007; for an example of such as study see Thomas, 2012).

In sum, with the current results in mind, the first take-home message is that psychotherapy researchers and clinical practitioners should be careful to interpret scores at the lower end of the scale when using unipolar screening instruments such as the BDI-II, for example to measure therapy effects. Perhaps, a different scale such as quality of life could be added to the set of instruments to ensure that the change is indeed positive. In general, clinicians should always be aware of the scale characteristics of the measurement instruments they purposefully select to monitor their patients' therapy progress or outcome in research and clinical practice. The second take-home message is that for persons with pretest-posttest

score differences that are close to statistical significant change (for the BDI-II raw score differences of 6 through 8), the change index that is used matters. It determines whether the same patient is classified as changed or unchanged. Therefore, we conclude that it is fruitful to embrace the advantages of the weighted metric θ and score-dependent reliability estimates of IRT modeling in the measurement of individual change.

Chapter 6
Epilogue

Through the work that is described in this thesis I showed that the development of high quality assessment scales in clinical psychology is a complicated endeavor. Clinical psychologists measure complex constructs, such as depression severity or interpersonal behavior, with a limited number of indicators that are strongly interrelated. Psychologists can interpret these complex constructs on different levels of the construct hierarchy and they also have to take into account various trade-offs such as those between test lengths and items inter-relatedness. Furthermore, when applying these scales in practice, there is the subjective nature of the self-report assessment and the different response biases that complicate clinical assessment even further. Notwithstanding these difficulties, psychologists have put much effort into the development of psychological assessment scales that measure many of the complex constructs that are used in clinical practice. In decades of research clinical psychologists have critically investigated and improved these scales with empirical studies in different samples and with different types of psychometric analyses.

In this thesis I have discussed and demonstrated how psychometric theories and methods, such as the bifactor model and the graded response model, can be used to gain new perspectives on the quality of assessment scales in clinical psychology. In the previous four chapters the results from four empirical studies on the DHS, the IIP-64, and the BDI-II in different clinical samples were reported, discussed, and related to clinical practice. I will now describe the main conclusions from the empirical results that relate to the three overarching issues I identified in the introduction chapter, followed by the overall insights, limitations, and recommendations for future research and clinical practice.

6.1 Discussion of Three Overarching Issues in this Thesis

6.1.1 Subscales in Clinical Practice

The first issue concerns the ongoing debate among researchers and clinical practitioners about the use and interpretation of subscale scores of screening instruments in clinical practice. From the results of the research presented I draw two main conclusions. The first conclusion is that IRT analyses once again confirmed that researchers and practitioners should be very careful when applying subscales scores as compared to total scale scores because measurement precision for subscale scores was lower than for total scale scores. Measurement precision for subscale scores was lower across all latent variable values, because the subscales consisted of fewer items than the total scales. Furthermore, some subscales were too heterogeneous (for example, the Vindictive, Domineering and Intrusive subscales of the IIP-64, see chapter 3) to allow for a reliable ordering of the scores according to the underlying latent variable. For all subscale scores discussed in the previous chapters, estimated reliability, as related to the scale information curves, was lower than .90 and in some cases lower than .80. Nunnally and Bernstein (1994) argued that a reliability lower than .90 is too low for individual decision making.

The second conclusion, based on the results from bifactor analyses for the DHS and BDI-II in Chapters 2 and 4, is that in order to be able to interpret subscale scores that are reliable and that explain additional subscale variance above the variance that is due to a general factor, a substantial number of items are needed that measure a *unique* aspect that the subscale is trying to measure. For example, it has been suggested in the literature that the items of the Pathways subscale of the DHS form a source of variance related to a unique underlying variable beyond the common variance that is accounted for by the general hope factor. Although, there was some unique variance explained by the Pathway subscale items, most variance was explained by the general hope factor. The results in Chapters 2 and 4 showed that it is difficult to make clinical interpretations based on subscale scores.

In general, it seems rather difficult for clinical psychologists to create subscales that really measure something different than the general construct and that are long enough to allow for reliable measurement. The obvious solution of creating longer subscales that are coherent and more distinct is not easy, because unlike for some measures in educational assessment the

number of items that can be generated in the clinical field to indicate a construct are limited. Furthermore, simply repeating the same item content over and over leads to a very homogenous scale that measures a smallband construct.

Emons, Sijtsma, and Meijer (2007) advised to create scales of at least 20 items to reach consistent individual classification results. In an educational context, Sinharay (2010; Sinharay et al., 2010) also showed that subscales should consist of at least 20 items and that they should be sufficiently distinct from other subscales (with disattenuated correlations less than .85) to have any hope of having added value[1]. Sinharay (2010) concluded that although "*several practitioners believe that subscores consisting of a few items may have added value if they are sufficiently distinct from each other* (…) *the results in this study provide evidence that is contrary to that belief. Subscores with 10 items were not of any added value even for a realistically extreme (low) disattenuated correlation of .70. The practical implication of this finding is that the test developers have to work hard (to make the subtests long and distinct) if they want subscores that have added value*" (p. 169). The findings in the different chapters in this thesis point to similar conclusions.

6.1.2 Factor Structure of Clinical Assessment Scales

The second issue concerns the factor structure of clinical scales and the measurement models that can be fit to the data of different constructs in clinical psychology. This issue is closely related to the use of subscale scores based on the bifactor results I just discussed. Often-used models to explain answering behavior on clinical scales are the one-factor model, correlated factors models, or hierarchically ordered factor models. Instead of using these models, I applied the bifactor model to investigate the factor structure of the hope and depression constructs as measured by the DHS and the BDI-II. The results showed that the bifactor models gave an adequate description of the data and that a general factor explained much of the subscale variance in these scales. It was interesting to compare different measurement models. For example, although factor loadings for separate factors in a correlated-trait factors

[1] A subscale score had added value when the subscale score provided more accurate diagnostic information (in the form of a lower mean squared error in estimating the true subscore) than the observed total score (Haberman, 2008).

model for the DHS and BDI-II were high, inspection of the corresponding group factors in a bifactor model showed that the magnitude of these loadings was much lower after controlling for the general factor. In fact, for almost all items the factor loadings of the general factor were higher than the loadings on the group factors. I conclude that when researchers choose a measurement model without a general factor, they may ignore to interpret the largest source of variance that underlies most items in a questionnaire.

In recent years several researchers have drawn this same conclusion for different types of clinical scales. For example, Ebesutani et al. (2011) recommended clinical practitioners to interpret only the full Negative Affectivity scale as opposed to lower order subfactors fear and distress in the Positive and Negative Affect Schedule for Children (PANAS-C). Simms, Grös, Watson, and O'Hara (2008) demonstrated that most of the Inventory of Depression and Anxiety Symptoms (IDAS) items were associated with a strong general distress factor. Reise et al. (2011) showed that for the Clinical Global Impression of Cognition in Schizophrenia (CGI-CogS) there was a large general psychiatric distress factor that accounted for 73% of the common variance. Also in the cognitive domain Canivez and Watkins (2010) stated firmly that clinical interpretations of intelligence should be made primarily at the level of general intelligence and not on a subscale level, because the general factor accounted for large proportion of the common variance (69.1%; see also Gignac, 2005, 2006, 2007, 2008). Thomas (2012) warned researchers that although different factors can be distinguished in the factor structure of the Brief Symptom Inventory, such a correlated trait factor structure '*can mislead researchers and clinicians into thinking that scales primarily measure distinct components of psychopathology. In fact, most items were most heavily influenced by the general psychiatric distress dimension*.' (p. 109). This does not imply that we should simply ignore other sources of variance. For example, the bifactor results from a study on the psychometric validity of the Psychiatric Diagnostic Screening Questionnaire (PDSQ) by Gibbons, Rush and Immekus (2009) provided evidence for the presence of a general psychiatric dimension as well as several relatively distinct diagnostic symptom sub-domains. However, the accumulated evidence suggests that although many clinical assessment scales include both broad and narrow factors, they should be interpreted at the broad level because a large general factor captures the largest amount of common variance in all items (see also Brunner et al., 2012; Emmons, 1995; Meijer, de Vries, & van Bruggen, 2011; Reise, 2012; Reise et al., 2010). Therefore, when their studies include clinical scales, researchers may

consider measurement models with a general factor in addition to other measurement models they investigate (Brunner et al., 2012; Gustafsson & Åberg-Bengtsson, 2010; Reise, Bonifay, & Haviland, 2012).

6.1.3 Measuring Individual Change with Clinical Assessment Scales

The third issue pertains to the measurement precision of scale scores in relation to individual change. First, we determined that the measurement precision of scale scores often varied across different score ranges. Because in many clinical scales items are selected to identify distress (or other clinical relevant constructs) these scales consist of items that indicate a certain amount of distress, whereas items that indicate more healthy symptoms are left out. In Chapters 3 and 5 it was demonstrated that the measurement precision of the IIP-64 subscale and BDI-II total scale was relatively large at the higher end of the scale, but low at the lower end of the scale. These scales function as they are intended to function. The scales can be used to discriminate persons with different levels of problematic interpersonal behavior (IIP-64) and different levels of depressive symptomatology (BDI-II), but they cannot be used to discriminate between different levels of "healthy" behavior. The IIP-64 subscales and the BDI-II total scale measure what Reise and Waller (2009) called a *quasi-trait* "a unipolar construct in which one end of the scale represents severity and the other pole represents its absence" (p. 31). Consequently, these scales provide the most test information or the smallest standard error for a limited range of scores.

Different authors found similar results for other clinical scales. For example, Reise and Haviland (2005) demonstrated that the items of the Cognitive Problems scale of the Minnesota Multiphasic Personality Inventory-2 were all located within a limited range of the trait. Doucette and Wolf (2009) showed that although the Life Status Questionnaire (LSQ) had a population reliability estimate for scale scores of .93, on an individual level more than 40% of the persons (those with low or mild scores) had a much lower measurement precision because test information was only high for a very limited range of scores (for more examples of clinical scales that measure quasi traits, see Cole et al., 2011; Emons et al., 2007; Meijer & Egberink. 2011; Purpura et al., 2010; Reise & Waller, 2009; Walters et al., 2011). Knowledge about the range of scores for which the items provide most test information, and knowledge about the score ranges where there are *item gaps* is important for clinical researchers and clinical practitioners.

Second, clinical scales, such as the BDI-II, that were originally constructed as screening instruments, are increasingly used for routine outcome monitoring and the measurement of individual change. In Chapter 5 we demonstrated that because measurement precision of scale scores may vary across scale scores it may also vary for pre- and posttest scores. We found that in a sample of clinical outpatients that completed the BDI-II twice (pre- and posttest scores), half of the posttest scores were on the low end of the depression scale where measurement precision was low. It is important that clinical psychologists realize that clinical scales that measure quasi-traits may have limited use for the measurement of change.

Third, the most often used method in clinical practice to determine whether change is statistically significant, the RCI index, assumes that the standard error is equal across measurements which may be not the case. Therefore, we used the Z-test, a change index based on IRT, which takes into account different standard errors for pre- and posttest score. Comparing results from the Z-test and the RCI we concluded that on average these change indices lead to similar results, but that for individuals for whom the change is located in the extreme score ranges or for whom change is close to statistically significant change the differences between these methods can lead to different classifications of change scores (deteriorated, unchanged, or improved).

6.2 Limitations and Future Research

There are several limitations of the studies conducted in this thesis that are important to mention. First, because the conclusions in this thesis are based on a limited number of clinical screening scales the generalizability of these conclusions to other clinical assessment scales should be made with care. Although other studies with clinical questionnaires seem to provide evidence for the same conclusions, future research may shed more light on the generalizability of the findings in this thesis (e.g., Reise et al., 2011). Future research of clinical scales in different samples should include information about the frequency distribution of the scores, and the measurement precision for those scores to determine whether they are suited for the population for which they are used. Researchers and clinical practitioners can either adjust instruments accordingly or add other instruments that measure the underrepresented end of the construct to the battery of questionnaires that they present to their patients.

Second, in the studies reported in this thesis I did not relate the general factor to external criteria. Thomas (2012) stated that '*bifactor models contribute greatly to the quantifying of general variance; however, defining the meaning of such dimensions is somewhat beyond the scope of mathematical modeling*' (p. 110). Different studies (e.g., Simms et al., 2008) have found positive correlations between the general factor scores of clinical scales with general psychiatric distress scale scores from various other questionnaires, that is, the distress and heightened negative affectivity that an individual experiences when applying for treatment. Future studies may investigate the meaning of the general factor for different clinical scales and differentiate between variance that is explained by, for example, distress and different response biases that influence self-report measurement.

Third, the results presented in the different chapters demonstrated some of the advantages that IRT offers to analyze clinical questionnaire data. A drawback of IRT analyses is that it requires larger sample sizes as compared to most CTT approaches (for example we need at least 500-1000 persons for GRM parameter estimation, see Kim & Cohen, 2002; Reise & Yu, 1990; Thissen et al., 2003) and the software that is available is sometimes not as user friendly as the computer programs many psychologists know by training such as SPSS. Furthermore, the latent variable metric is different from the raw score metric and the (normalized) t-score metric that are often used in practice (e.g., de Beurs, 2010). In future research the performance of these different scores can be compared for different scales and samples. Many researchers (e.g., Embretson & Reise, 2000) have underlined that using the latent variable metric has important advantages over the CTT metric, such as individual tailored standard errors, higher measurement precision for extreme scores, interval measurement, and weighted subscale scores. Sinharay (2010) demonstrated that weighted subscale scores have a higher added value as compared to unweighted scores. Cole et al. (2011) showed that '*using IRT-derived information about symptom severity and discriminability substantially enhanced precision in the measurement of depression severity*' (p. 827) as compared to using the unweighted total scores and reliability coefficients.

A fourth limitation is that in the study reported in Chapter 5 it was not possible to provide easy guidelines for clinicians to help to understand when change from pre- to posttest scores can be interpreted as reliable change. Based on findings in this study, I concluded that measurement precision differed across some score ranges, and that for screening

questionnaire posttest scores are typically less reliable as compared to the pretest scores. Doucette and Wolf (2009) stated that this '*might inadvertently lead to the interpretation of stable scores as a lack of psychotherapy effectiveness when in reality, it is likely a consequence of the measure: insufficient item coverage at the ability location for those persons*' (p. 386) who report few or no problems (see also Reise & Havliand, 2005). The exact consequences for clinical practice and research are unknown and should be further investigated. It may be interesting to investigate to what degree the lack of measurement precision on unipolar scales for one end of the continuum influences the outcome effects that are reported in therapy effect studies.

One final remark. Although Nunnally and Bernstein (1994) recommended that a reliability of .90 is at least required for scores that are used to make individual predictions or decisions, it should be noted that in clinical practice a score from a screening questionnaire is almost always combined with other sources of information, as I described in the case of the 22-year old student in the introduction of this thesis. This increases the overall reliability of the evidence that is used to substantiate clinical decision making. For example, the measurement precision for the subscales of the IIP-64 for the three elevated scale scores for the 22-year old student were rather high. This information can be combined with information from the other (opposite) subscale scores that were not elevated. The outcome of the IIP-64 indicated interpersonal distress due to a dependent and submissive attitude and not due to the opposite, a cold attitude. Clinical psychologists are trained to combine scale score with information from an interview and their observations. Future researchers should investigate whether, such as Sinharay et al. (2010) mentioned '*combining some subscores may result in subscores that have higher reliability and hence added value*' (p. 570) so that the research more accurately reflects the actual procedure of combining sources of information in assessment in clinical practice.

6.3 Recommendations for Clinical Practice

When psychologists interpret subscale scores as reflecting unique and different constructs, they often do not take into account that (1) these scores may share much more variance with other subscale scores than they contribute to the unique part of the subscale variance, and (2) that measurement precision for subscale scores is relatively low and confidence intervals are

generally large. For clinical practitioners the take-home message is that in many cases it is wise to mainly consider the total scale scores of psychological screening instruments rather than to interpret subscale scores. As a clinician I am aware of the potentional diagnostic value of combining subscale scores. When combining subscales scores, however, it is very important to consider the psychometric quality of the subscales. For example, as explained for the IIP-64, the psychometric quality of the Vindictive, Domineering, and Intrusive subscales is too low to be of practical value. The other subscales can be interpreted carefully in combination with other observations after first having interpreted the total scale score in relation to the personality and psychopathology of the patient at hand.

Furthermore, on an institutional and policy level the take-home message is that IRT techniques can be useful tools in different clinical test applications. However, IRT is seldom used in clinical practice. The clinical field should utilize the advantages of modern psychometric approaches more often. Besides the topics I discussed in the foregoing chapters another interesting application is computer-adaptive testing. Instruments can be developed that adapt to the latent variable level of the patient to achieve a desirable level of measurement precision with a fewer number of items than for linear tests. An interesting example can be found in the open-source project the Patient Reported Outcomes Measurement Information System (PROMIS) project (Cella et al., 2010; Gerson et al., 2010; Liu et al., 2010; Rothrock et al., 2010). PROMIS contains different series of item banks that measure different homogeneous concepts that can be administered with computer adaptive testing (see http://www.nihpromis.org/ and for a Dutch version see http://www.kmin-vumc.nl/promis_13_0.html).

Having said all this, I hope to have shown, that almost sixty years after Cronbach's (1954) call to Psychometrikans to undertake a journey to the planet of Clinicia, psychometric missions to Clinicia are still worthwhile to undertake. I challenge psychometricians to increase their efforts to show how their research findings can be applied to everyday clinical practice. And I challenge clinical psychologist to increase their efforts to utilize the new methods provided by modern psychometrics. There are many opportunities for shared research projects. Moreover, modern psychometric techniques can be implemented in clinical assessment procedures to improve the quality of assessment in the daily practice of clinical psychology.

References

Acton, G. S., & Revelle, W. (2004). Evaluation of ten psychometric criteria for circumplex structure. *Methods of Psychological Research Online*, 9-1.

Acton, G.S., Revelle, W. (2002). Interpersonal personality measures show circumplex structure based on new psychometric criteria. *Journal of Personality Assessment*, 79, 446-471.

Alden, L.E., Wiggins, J.S., & Pincus, A.L. (1990). Construction of circumplex scales for the assessment of Interpersonal Problems. *Journal of Personality Assessment*, 55, 521–536.

Al-Turkait, F. A., & Ohaeri, J. U. (2010). Dimensional and hierarchical models of depression using the Beck Depression Inventory-II in an Arab college student sample. *BMC Psychiatry*, 10, 60-74.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (4th ed-TR)*. Washington, DC: Task Force.

Arnau, R. C., Meagher, M. W., Norris, M. P., & Bramson, R. (2001). Psychometric evaluation of the Beck Depression Inventory–II with primary care medical patients. *Health Psychology*, 20, 112–119.

Arnau, R.C., Rosen, D.H., Finch, J.F., Rhudy, J.L., & Fortunato, V.J. (2007). Longitudinal Effects of Hope on Depression and Anxiety: A Latent Variable Analysis. *Journal of Personality*, 75 , 43–64.

Arrindell, W. A., & Ettema, J. (1986). *SCL-90: een multidimensionele psychopathologie-indicator* [SCL 90: Manual for a Multidimensional Indicator of Psychopathology]. Lisse, the Netherlands, Swets & Zeitlinger.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, 16, 397- 438.

Atkins, D.C., Bedics, J.D., McGlinchey, J.B., & Beauchaine, T.P. (2005). Assessing clinical significance: Does it matter which method we use? *Journal of Consulting and Clinical Psychology*, 73, 982–989.

Babyak, M. A. Snyder, C.F., Yoshinobu, L. (1993). Psychometric properties of the Hope scale: A confirmatory factor analysis. *Journal of Research in Personality*, 27, 154-169.

Barkham, M., Hardy, G.E., & Startup, M. (1996). The IIP-32: development of a short version of the Inventory of Interpersonal Problems. *British Journal of Clinical Psychology*, 35, 21-35.

Bateman, A. & Fonagy, P. (2004) *Psychotherapy for Borderline Personality Disorder: Mentalisation Based Treatment.* Oxford: Oxford University Press.

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, 82, 60-70

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory–II*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment,* 67, 588–597.

Beck, A. T., Steer, R. A., Brown, G. K., & van der Does, A. J. W. (2002). *BDI–II–NL Handleiding* [BDI–II–Dutch manual]. Lisse, the Netherlands, Psychological Corporation.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, 74, 137–143.

Berghout, C.C., Zevalkink, J. & de Jong, J.T. (2011). Symptomatic distress and personality functioning before, during, and after long-term psychoanalytic treatment. Journal of the American Psychoanalytic Association, 59 (3), 583-588.

Beurs, E. de., Barendreft, M., Flens, G., Dijk, E. van, Huijbrechts, I, & Meerding, W. J. (2012). Vooruitgang in de behandeling meten. Een vergelijking van vragenlijsten voor zelfrapportage. [Measuring progress in treatment. A comparison between self-report questionnaires]. *Maandblad Geestelijke Volksgezondheid*, 67, 259–264.

Binder, P., Holgersen, H., & Nielsen, G. H. (2010). What is a “good outcome” in psychotherapy? A qualitative exploration of former patients' point of view. *Psychotherapy Research*, 20:3, 285-294.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.

Brouwer D., Meijer, R. R., Weekers, A. M., & Baneke, J. J. (2008). On the dimensionality of the Dispositional Hope Scale, *Psychological Assessment*, 20, 310-315.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2012, July 16). On the Factor Structure of the Beck Depression Inventory–II: G Is the Key. *Psychological Assessment*. Advance online publication. doi: 10.1037/a0029228

Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika*, 57, 469–497.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*. Advance online publication. doi:10.1111/j.1467-6494.2011.00749.x

Buckley, T. C., Parker, J. D., & Heggie, J. (2001). A psychometric evaluation of the BDI–II in treatment seeking substance abusers. *Journal of Substance Abuse Treatment*, 20, 197–204.

Cai, L., Thissen, D., & du Toit, S. (2011a). *IRTPRO 2.1 for Windows*. Chicago: Scientific Software International.

Cai, L., Yang, J. S., & Hansen, M. (2011b). Generalized full-information item bifactor analyses. *Psychological Methods*, 16, 221-248.

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): Exploratory and higher order factor analyses. *Psychological assessment*, 22(4), 827.

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research*, 30, 429–452.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D. J., Choi, S. W., Cook, K. F., DeVellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E., Pilkonis, P., Revicki, D., Rose, M., Weinfurt, K., & Hays, R. D. on behalf of the PROMIS Cooperative Group. (2010). Initial item banks and first wave testing of the Patient–Reported Outcomes Measurement Information System (PROMIS) network: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179-94.

Chang, A. C. (2003). A Critical Appraisal and Extension of Hope Theory in Middle-Aged Men and Women: Is it Important to Distinguish Agency and Pathways Components? *Journal of Social and Clinical Psychology*, 22, 121-143.

Chen, F. F., West S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41, 189-225.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy*, 15, 305–308.

Clarkin, J. F., Yeomans, F., & Kernberg, O. F. (2006). *Psychotherapy of borderline personality: Focusing on object relations*. Arlington, VA: American Psychiatric Publishing.

Cole, D. A., Cai, L., Martin, N. C., Findling, R. L., Youngstrom, E. A., Garber, J., & ... Forehand, R. (2011). Structure and measurement of depression in youths: Applying item response theory to clinical data. *Psychological Assessment*, 23(4), 819-833. doi:10.1037/a0023518.

Collins, L. M. (1996). Is Reliability Obsolete? A Commentary on 'Are Simple Gain Scores Obsolete,'. *Applied Psychological Measurement*, 20, 289-92.

Cook, K. F., Kallen, M. A., & Amtmann D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research*, 18, 447-460.

Cronbach, L. J. (1954). Report on a Psychometric Mission to Clinicia. *Psychometrika*. 19, 263-270.

Cronbach, L. J., Furby, L. (1970). How We Should Measure 'Change' – Or Should We?. *Psychological Bulletin*, 74, 68-80.

de Beurs, E. (2010). De genormaliseerde T-score. Een 'euro' voor testuitslagen. [The normalized T-score]. *Maandblad geestelijke gezondheidszorg*, 65(9) 684-695.

Derogatis, L. R. (1983). *SCL-90-R, administration, scoring, and procedures manual (2nd edition)*. MD: Clinical Psychometric Research.

Doucette, A., & Wolf, A.W. (2009). Questioning the measurement precision of psychotherapy research. *Psychotherapy Research*, 19, 374-389.

Dozois, D. J. A., Dobson, K. S., & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory–II. *Psychological Assessment*, 10, 83–89.

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment*, 20, 55-62.

Ebesutani, C., Smith, A., Bernstein, A., Chorpita, B. F., Higa-McMillan, C., & Nakamura, B. (2011). A bifactor model of negative affectivity: Fear and distress components among younger and older youth. *Psychological assessment*, 23(3), 679.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ. Erlbaum.

Emmons, R. A. (1995). Levels and domains in personality – an introduction. *Journal of Personality*, 63, 341-364.

Emons, W. H. M., Meijer, R. R. ,Denollet, J. (2007). Negative affectivity and social inhibition in cardiovascular disease: evaluating Type D personality and its assessment using item response theory. *Journal of Psychosomatic Research*, 63, 27–39.

Emons, W., Sijtsma, K., & Meijer, R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, 12, 105-120.

Finkelman, M., Weiss, D., & Kim-Kang, G. (2010). Item Selection and Hypothesis Testing for the Adaptive Measurement of Change. *Applied Psychological Measurement*, 34(4), 238-254.

Gershon, R. C., Rothrock, N. E., Hanrahan, R. T., Jansky, L. J., Harniss, M., Riley, W. (2010). The development of a clinical outcomes survey research application: Assessment Center. *Quality of Life Research*, 19(5), 677-85.

Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of Psychiatric Research*, 43, 401-410.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika*, 57, 53-61.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D., Segawa, E., Bhaumik, D. K., Kupfer, D., Frank, E., Grochocinski, V., & Stover, A. (2007). Full-Information Item Bi-Factor Analysis of Graded Response Data. *Applied Psychological Measurement*, 31, 4-19.

Gignac, G. E. (2005). Revisiting the factor structure of the WAIS–R: Insights through nested factor modeling. *Assessment*, 12, 320–329.

Gignac, G. E. (2006). The WAIS–III as a nested factors model: A useful alternative to the more conventional oblique and higher-order models. *Journal of Individual Differences*, 27, 73–86.

Gignac, G. E. (2007). Multi-factor modeling in individual differences research: Some recommendations and suggestions. *Personality and Individual Differences*, 42, 37-48.

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor?. *Psychology Science*, 50(1), 21.

Grosse-Holtforth, M., Lutz, W., & Grawe, K. (2006). Structure and change of the IIP-D pre- and postpsychotherapy. *European Journal of Psychological Assessment*, 22, 98–103.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored Internet tests: The z-test and the likelihood ratio test. *International Journal of Selection and Assessment*, 18, 351-364.

Gurtman, M. B. (1996). Interpersonal problems and the psychotherapy context: the construct validity of the inventory of interpersonal problems. *Psychological Assessment*, 8, 241-255.

Gustafsson, J. E., & Åberg-Bengtsson, L. (2010). *Unidimensionality and the interpretability of psychological instruments*. In S. E. Embretson (Ed.), Measuring psychological constructs (pp. 97–121). Washington, DC: American Psychological Association.

Haberman, S. J. (2008).When can subscores have value? Journal of Educational and *Behavioral Statistics*, 33, 204–229.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item Response Theory and Health Outcomes Measurement in the 21st Century. *Medical Care*, 38-9 supplement II, II-28 – II-42.

Hiller, W., Schindler, A. C., & Lambert, M. J. (2012): Defining response and remission in psychotherapy research: A comparison of the RCI and the method of percent improvement. *Psychotherapy Research*, 22:1, 1-11.

Holzinger, K. J., & Swineford, F. (1937). The bifactor method. *Psychometrica*, 2, 41-54.

Horowitz, L. M., Alden, L. E., Wiggins, J. S., & Pincus, A. L. (2000). *Inventory of Interpersonal Problems: Manual.* New York: The Psychological Corporation Harcourt.

Horowitz, L. M., Rosenberg, S. R. & Bartholomew, K. (1993). Interpersonal problems, attachment styles, and outcome in brief dynamic psychotherapy. *Journal of Consulting and Clinical Psychology*, 61, 549-560.

Horowitz, L.M., & Vitkus, J. (1986). The interpersonal basis of psychiatric symptoms. *Clinical Psychology Review*, 6, 443-469.

Horowitz, L.M., Rosenberg, S.E., Baer, B.A, Ureño, G., & Villaseñor, V.S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology*, 56, 885-892.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to unparameterized model misspecification. *Psychological Methods*, 3, 424–453.

Huber, D., Henrich, G., & Klug, G. (2007). The Inventory of Interpersonal Problems (IIP): Sensitivity to change. *Psychotherapy Research*, 17, 474-481.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12–19.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, 15, 336–352.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, 67, 300–307.

Kelderman, H. (1997). *Loglinear multidimensional item response models for polytomously scored items*. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 85-100). New York: Springer.

Kendall, P. C., Marrs-Garcia, A., Nath, S., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, 67, 285–299.

Kim, S. & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, 26 (1), 25- 41.

Lambert, M. J. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research*, 17, 1-14.

Lambert, M. J., & Ogles, B. M. (2009). Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data. *Psychotherapy Research*, 19:4-5, 493-501.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A.,Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice*, 10, 288 –301.

Leising, D., Rehbein, D., & Eckardt, J. (2009). The Inventory of Interpersonal Problems (IIP-64) as a screening measure for Avoidant Personality Disorder. European Journal of *Psychological Assessment*, 25, 16-22.

Leising, D., Rehbein, D., & Sporberg, D. (2007). Validity of the Inventory of Interpersonal Problems (IIP-64) for Predicting Assertiveness in Role-Play Situations. *Journal of Personality Assessment*, 89, 116-125.

Liu, H. H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., & Hays, R. D. (2010). Representativeness of the PROMIS Internet panel. *Journal of Clinical Epidemiology*, 63(11), 1169-78.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Maassen, G. H. (2001). The unreliable change of reliable change indices. *Behaviour Research and Therapy*, 39, 495–498.

Magaletta, P. R., & Oliver, J. M. (1999). The hope construct, will and ways: Their relations with self-efficacy, optimism, and general well-being. *Journal of Clinical Psychology*, 55, 539 –551.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519–530.

McCullough, L., Kuhn, N., Andrews, S., Kaplan, A., Wolf, J., & Hurley, C. (2003). *Treating Affect Phobia: a Manual for Short-Term Dynamic Psychotherapy*, New York: Guilford Press.

McGlinchey, J. B., Atkins, D. C., & Jacobson, N. S. (2002). Clinical significant methods: Which one to use and how useful are they? *Behavior Therapy*, 33, 529–550.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory. *Psychological Methods*, 9, 354 –368.

Meijer, R. R., & Egberink, I. J. L. (2011). An Item Response Theory Analysis of Harter's Self-Perception Profile for Children or Why Strong Clinical Scales Should be Distrusted. *Assessment*, 2, 201-212.

Meijer, R. R., de Vries, R. M., & van Bruggen, V. (2011). An evaluation of the Brief Symptom Inventory-18 using item response theory: Which items are most strongly related to psychological distress? *Psychological Assessment*, 23, 193-202.

Meijer, R.R., Egberink, I.J.L., Emons, W.H.M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using Item Response Theory: An illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment*, 90, 227-238.

Mellenbergh, G. J., & Van den Brink, W. P. (1998). The measurement of individual change. *Pschychological Methods*, 3-4, 470-485.

Mellenbergh, G.J. (1999). A note on simple gain score precision. *Applied Psychological Measurement*, 23, 87-89.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. New York/Berlin: De Gruyter.

Molenaar, I.W. & Sijtsma, K. (2000). *MSP5.0 for windows. User's manual*. Groningen, The Netherlands: ProGamma.

Muthén, L. K., & Muthén, B. O. (1998 –2006). *MPlus user's guide (4th ed.)*. Los Angeles, CA: Muthén & Muthén.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3, 1966, 1-18.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory (3rd Edition)*. McGraw-Hill Series in Psychology, McGraw-Hill, Inc., New York: NY, 264-265.

Ogles, B. M., Lunnen, K. M., & Bonesteel, K. (2001). Clinical significance: History, application and current practice. *Clinical Psychology Review*, 21, 421–446.

Osman, A., Barrios, F. X., Gutierrez, P. M., Williams, J. E., & Bailey, J. (2008). Psychometric Properties of the Beck Depression Inventory-II in Nonclinical Adolescent Samples. *Journal of Psychopathology and Behavioral Assessment*, 64, 83-102.

Osman, A., Downs, W. R., Barrios, F. X., Kopper, B. A., Guttierrez, P. M., & Chiros, C. E. (1997). Factor structure and psychometric characteristics of the Beck Depression Inventory–II. *Journal of Psychopathology and Behavioral Assessment*, 19, 359–376.

Percevic, R., Lambert, M. J., & Kordy, H. (2006). What is the predictive value of responses to psychotherapy for its future course? Empirical explorations and consequences for outcome monitoring. *Psychotherapy Research*, 16, 364-373.

Pincus, A.L., Gurtman, M. B., & Ruiz, M. A. (1998). Structural Analysis of Social Behavior (SASB): Circumplex Analyses and Structural Relations With the Interpersonal Circle and the Five-Factor Model of Personality. *Journal of Personality and Social Psychology*, 74, 1629-1645.

Purpura, D. J., Wilson, S. B., & Lonigan, C. J. (2010). Attention-deficit/hyperactivity disorder symptoms in preschool children: Examining psychometric properties using item response theory. *Psychological Assessment*, 22(3), 546-558. doi:10.1037/a0019581

Purpura, D. J., Wilson, S. B., & Lonigan, C. J. (2010). Attention-deficit/hyperactivity disorder symptoms in preschool children: Examining psychometric properties using item response theory. *Psychological assessment*, 22(3), 546-558.

Puschner, B., Kraft, S., & Bauer, S. (2004). Interpersonal Problems and Outcome in Outpatient Psychotherapy: Findings From a Long-Term Longitudinal Study in Germany. *Journal of Personality Assessment*, 83(3), 223-234.

Quilty, L. C., Zhang, A. Z., & Bagby, R. M. (2010). The latent symptom structure of the Beck Depression Inventory–II in outpatients with major depression. *Psychological Assessment*, 22, 603-608.

R Development Core Team (2010). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org/.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.

Reise, S. P. (2012). The rebirth of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2012) Scoring and Modeling Psychological Measures in the Presence of Multidimensionality. *Journal of Personality Assessment*, DOI:10.1080/00223891.2012.725437

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, 84, 228–238.

Reise, S. P., & Henson, J. (2003). A Discussion of Modern Versus Traditional Psychometrics As Applied to Personality Assessment Scales. *Journal of Personality Assessment*, 81, 93-103.

Reise, S. P., Moore, T. N., & Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. Journal of *Personality Assessment*, 96, 544-559.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcome measures. *Quality of Life Research*, 16, 19 –31.

Reise, S.P., Scheines, R., Widaman, K.F., & Havilund, M. G. (2012). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. Manuscript submitted for publication.

Reise, S. P., Ventura, J., Keefe, R. S., Baade, L. E., Gold, J. M., Green, M. F., ... & Bilder, R. (2011). Bifactor and item response theory analyses of interviewer report scales of cognitive impairment in schizophrenia. *Psychological assessment*, 23(1), 245-261.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5, 27–48.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, 12, 287–297.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, 27 (2), 133—144.

Revelle, W. (2012). *Psych: Procedures for Psychological, Psychometric, and Personality Research. R package Version 1.2.1.* Retrieved from http://personality-project.org/r.

Roesch, S. C., & Vaughn, A. A. (2006). Evidence for the factorial validity of the Dispositional Hope Scale, cross-ethnic and cross-gender measurement equivalence. European *Journal of Psychological Assessment*, 22, 78 – 84.

Rothrock, N. E., Hays, R. D., Spritzer, K., Yount, S. E., Riley, W., and Cella, D. (2010). Relative to the general US population, chronic diseases are associated with poorer health–related quality of life as measured by the Patient–Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology*, 63(11), 1195-204.

Ruiz, M. A., Pincus, A.L., Borkovec, T. D., Echemendia, R.J., Castonguay, L.G., & Ragusea, S.A. (2004). Validity of the Inventory of Interpersonal Problems for Predicting Treatment Outcome: An Investigation With The Pennsylvania Research Network. *Journal of Personality Assessment*, 83, 213-222.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34, 100.

Samejima, F. (1997). *Graded reponse model*. In W.J. van der Linden & R.K. Hambleton (Eds.), Handbook of modern item response theory (pp. 85-100). New York: Springer-Verlag.

Santor, D. A., Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of Item Response models. *Psychological Assessment*, 10, 345–59.

Schmid, J., & Leiman, J. N. (1957). The development of hierarchical factor solutions. Psychometrika, 22, 53–61.

Seligman, M. E. P. (2005). *Positive psychology, positive prevention, and positive therapy.* In C. R. Snyder and S. J. Lopez (Eds.), Handbook of positive psychology (pp. 3–9). London: Oxford University Press.

Sijtsma, K. (2012). Future of Psychometrics: Ask what Psychometrics can do for Psychology. *Psychometrika*, 77(1), 4-20.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Simms, L. J., Grös, D. F., Watson, D., & O'hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety*, 7, 34–46.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47(2), 150-174.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, 45, 553-573.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting Diagnostic Scores in Educational Testing: Temptations, Pitfalls, and Some Solutions. *Multivariate Behavioral Research*, 45(3), 553-573.

Snyder, C. R. (2000). *Handbook of hope: Theory, measures, and applications*. San Diego, CA: Academic Press.

Snyder, C. R. (2004). Hope and depression: A light in the darkness. *Journal of Social and Clinical Psychology*, 23, 347–351.

Snyder, C. R., Harris, C., Anderson, J. R., Holleran, S. A., Irving, L. M., Sigmon, S. T., Yoshinobu, L., Gibb, J., Langelle, C., & Harney, P. (1991). The will and the ways: Development and validation of an individual differences measure of hope. *Journal of Personality and Social Psychology*, 60, 570-585.

Soltz, S., Budman, S., Demby, A., & Merry, J. (1995). A Short Form of the Inventory of Interpersonal Problems Circumplex Scales. *Psychological Assessment*, 2, 53-63.

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology*, 63, 1044 –1048.

Spielberger, C. D. (1983) *Manual for the State-Trait Anxiety Inventory (STAI-form Y).* Palo Alto, CA: Consulting Psychologists Press.

Steer, R. A., Ball, R., Ranieri, W. F., & Beck, A. T. (1999). Dimensions of the Beck Depression Inventory–II in clinically depressed outpatients. *Journal of Clinical Psychology*, 55, 117–128.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, 69, 613–625.

Thissen, D., Chen, W.H., & Bock, R.D. (2003) *MULTILOG (version 7).* [computer software]. Lincolnwood (Ill): Scientific Software International.

Thomas, M. L. (2011a). *Modern psychometric theory in clinical assessment* (Doctoral dissertation, Arizona State University).

Thomas, M. L. (2011b). The value of item response theory in clinical assessment: A review. *Assessment*, 18(3), 291–307.

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological assessment*, 24-1, 101-113.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology*, 112, 578-598.

Tracey, T., Rounds, J., & Gurtman, M. (1996). Examination of the General Factor with the Interpersonal Circumplex Structure: Application to the Inventory of Interpersonal Problems. *Multivariate Behavioral Research*, 31, 441.

Valkonen, J., Hänninen, V., & Lindfors, O. (2011). Outcomes of psychotherapy from the perspective of the users, *Psychotherapy Research*, 21-2, 227-240.

Van der Ploeg, H. M. (2000). *Handleiding bij de zelfbeoordelingsvragenlijst, een Nederlandstalige bewerking van de Spielberger State-Trait Anxiety Inventory* [Manual of the Self-Assessment Questionnaire (SAQ). A Dutch revision of the Spielberger State–Trait Anxiety Inventory (STAI-DY)]. Lisse, the Netherlands: Swets Test Publishers.

Vanheule, S., Desmet, M., & Rosseel, Y. (2006). The factorial structure of the Dutch translation of the Inventory of Interpersonal Problems: A test of the long and short versions. *Psychological Assessment*, 18, 112 117.

Vanheule, S., Desmet, M., Groenvynck, H., Rosseel, Y., & Fontaine, J. (2008). The factor structure of the Beck Depression Inventory–II. *Assessment*, 15, 177–187.

Viljoen, J. L., Grant, L. I., Griffiths, S., & Woodward, T. S. (2003). Factor structure of the Beck Depression Inventory–II in a medical outpatient sample. *Journal of Clinical Psychology in Medical Settings*, 10, 289–291.

Vittengl, J. R., Clark, L. A., & Jarrett, R. B. (2003). Interpersonal problems, personality pathology, and social adjustment after cognitive therapy for depression. *Psychological Assessment*, 15, 29-40.

Walters, G. D., Hagman, B. T., Cohn, A. M. (2011). Toward a hierarchical model of criminal thinking: Evidence from item response theory and confirmatory factor analysis. *Psychological Assessment*, 23(4), 925-936.

Ward, L. C. (2006). Comparison of factor structure models for the Beck Depression Inventory -II. *Psychological Assessment*, 18, 81-88.

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children–fourth Edition among a national sample of referred students. *Psychological Assessment*, 22, 782–787.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, 20, 59–69.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment*, 82, 50-59.

Wu, J., King, K. M., Witkiewitz, K., Racz, S. J., & McMahon, R. J. (2012). Item analysis and differential item functioning of a brief conduct problem screen. *Psychological Assessment*, 24(2), 444-454.

Yung, Y., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64, 113-128.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's alpha, Revelle's beta, and McDonald's omega h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70, 123-133.

Summary

The overarching aim of this thesis is to investigate the usefulness of modern psychometric approaches in clinical practice. The bifactor model is used to investigate the factor structure of two clinical screening questionnaires and the degree to which subscales scores of these questionnaires can be used to reliably interpret specific content areas. Furthermore, the graded response model (GRM) is used to describe conditions under which scores on clinical screening instruments are reliable indicators of individual change in psychological functioning. Chapter 1 starts with a description of how psychologists use screening questionnaires in clinical practice. Also, the overarching issues are further explained and the psychometric models are introduced.

In Chapter 2, the problem under investigation is the dimensionality of the Dispositional Hope Scale (DHS). Because researchers have made different recommendations with regard to the dimensionality of the DHS, it is unclear whether the use of subscale scores can be defended. We compared the analyses of a one-factor model, a two-factor model and a bifactor model in three samples: a student sample, a sample of psychiatric inpatients, and a sample of delinquents. The results indicated that the best choice is to consider the scale as a unidimensional scale. The items measure one construct and there is very little unique variance that is explained by the Pathways or Agency items above the general factor Hope.

In Chapter 3, the psychometric quality of the eight subscales of the Inventory of Interpersonal Problems 64 (IIP-64) is evaluated in a large sample of clinical outpatients. It is unclear how well the IIP-64 subscales tap the entire range of the underlying interpersonal problems dimensions. We used results from different IRT analyses to investigate the reliability of subscale scores for different ranges of scale scores. The results showed that five of the eight IIP-64 subscales (Cold, Socially Avoidant, Nonassertive, Exploitable, Overly Nurturant) formed scales of medium quality and that for three subscales (Vindictive, Domineering and Intrusive) the items were unscalable. Measurement precision differed across the latent trait ranges for all scales. The main conclusion is that when using IIP-64 subscales scores in clinical practice clinicians should interpret these scores with care because items do not tap the entire range of severity and three subscales do not allow precise measurement.

In Chapters 4 and 5 two studies are presented on the psychometric quality of the Beck Depression Inventory II (BDI-II). In Chapter 4, the problem is presented through a discussion in the research literature about the dimensionality of the BDI-II. Bifactor analyses are used to answer the question whether BDI-II data are unidimensional enough to scale persons according to their total depression score or whether the use of subscales scores should be preferred. We compared results from a one-factor model and different two-factor, three-factor, and bifactor models in a large sample of clinical outpatients. We observed that although total scale score variation reflected multiple sources of variance due to clustered item content, differences in factor loadings between the unidimensional model and the general factor from the bifactor models were small and the general factor explained by far the most variance in the bifactor models. Consequently, we concluded that the presence of multidimensionality did not handicap our ability to interpret the BDI-II as one scale. In fact, based on the results we concluded that clustering of items into separate dimensions and consequently scoring of subscales could hardly be justified. There is more common variance to the BDI-II factors than unique variance which implies that clinical practitioners should be careful when interpreting subscale scores, because these subscale scores are highly related to the general construct.

In Chapter 5, first we analyzed BDI-II change data from a sample of clinical outpatients with pre- and post treatment scores to determine the conditions under which scores on the BDI-II are reliable indicators of individual change in depression. Results showed that for the mild, moderate, and severe depression categories the measurement precision was high, but that measurement was inaccurate for low depression levels. Furthermore, the results demonstrated that different unweighted BDI-II total scores can indicate the same latent depression level and that the raw score metric is not suited for interval measurement. Second, we compared an IRT-based change index, the Z-test, to the often-used reliable change index (RCI). Results showed that on a group level there were no big differences in outcomes between the RCI and Z-test, but that on an individual level, RCI and Z-test outcomes lead to different interpretations of individual change.

Finally, in the epilogue I compared these research findings with results from other studies and tried to sketch a more general picture of the impact of the studies. Furthermore, I discussed some limitations of the studies in this thesis. I argued that results from modern psychometric

approaches, such as the GRM and the bifactor analyses, can guide the continued evaluation and improvement of assessment in the field of clinical psychology.

Samenvatting (Summary in Dutch)

Dit proefschrift beschrijft een viertal studies waarin centraal staat hoe moderne psychometrische methoden van waarde kunnen zijn voor de dagelijkse praktijk van klinisch psychologen. Met het bifactor model onderzoeken we de factorstructuur van twee klinische vragenlijsten en de mate waarin psychologen scores van subschalen op betrouwbare wijze kunnen gebruiken bij het interpreteren van specifieke deelgebieden van deze vragenlijsten. Ook beschrijven we door middel van het graded response model (GRM) de voorwaarden waaronder scores op klinische vragenlijsten een betrouwbare indicatie kunnen zijn van individuele verandering in psychologisch functioneren. In Hoofdstuk 1 beschrijf ik het gebruik van klinische vragenlijsten in de praktijk, benoem ik drie belangrijke onderzoekslijnen en introduceer ik de psychometrische modellen die in dit proefschrift worden gebruikt.

In Hoofdstuk 2 staat de dimensionaliteit van de Dispositional Hope Scale (DHS) centraal. Omdat onderzoekers allerlei verschillende aanbevelingen hebben gedaan met betrekking tot de dimensionaliteit van DHS is het onduidelijk of het gebruik van subschaalscores kan worden verdedigd. We vergeleken de resultaten van een een-factor model, een twee-factor model en een bifactor model in drie steekproeven: studenten, psychiatrische patiënten, en delinquenten. De resultaten laten zien dat de DHS het beste als een eendimensionale schaal kan worden beschouwd. Met de items van de DHS wordt één begrip gemeten, Hoop. Er is weinig unieke variantie die nog door de subschalen Pathways en Agency wordt verklaard wannneer de gemeenschappelijke variantie al verklaard is door een algemene Hoop factor.

In Hoofdstuk 3 evalueren we de psychometrische kwaliteit van acht subschalen van de Inventory of Interpersonal Problems 64 (IIP-64) in een grote steekproef van patiënten die ambulant psychologische zorg ontvangen. Wij gebruikten resultaten van verschillende item reponse theorie analyses om de betrouwbaarheid van subschaalscores voor verschillende niveaus van de te meten eigenschap te onderzoeken. De resultaten toonden aan dat vijf van acht IIP-64 subschalen (Afstandelijk, Sociaal geremd, Onderworpen, Aanpassend, Zelf-opofferend) van middelmatige kwaliteit zijn en dat voor drie subschalen (Zelfgericht, Controlerend en Behoeftig) de items onschaalbaar zijn. Ook vonden we dat voor alle schalen

de meetnauwkeurigheid afhankelijk was van de hoogte van de score. De belangrijkste conclusie is dat psychologen die IIP-64 subschaalscores gebruiken in hun dagelijkse praktijk zich moeten realiseren dat voor sommige scores de meetfout erg groot is. Dit geldt voor alle scores op de drie subschalen die van zeer lage kwaliteit waren en in het algemeen voor lage scores op alle subschalen.

In Hoofdstuk 4 en Hoofdfstuk 5 worden twee studies beschreven over de psychometrische kwaliteit van Beck Depression Inventory II (BDI-II). In Hoofdstuk 4 bespreken we eerst de verschillende bevindingen en conclusies over de dimensionaliteit van BDI-II in de recente onderzoeksliteratuur. Door middel van bifactor analyse onderzoeken we of de BDI-II eendimensionaal genoeg is om totaalscores te gebruiken of dat de BDI-II moet worden beschouwd als een multidimensionele vragenlijst en dat subschaalscores kunnen worden gebruikt. Wij vergeleken de resultaten van een-factor model, verschillende twee en drie-factor modellen en een aantal bifactor modellen in een grote steekproef van poliklinische patiënten. Hoewel verschillende itemclusters enige unieke variantie verklaarden, vonden we dat de verschillen in factorenladingen tussen het eendimensionale model en de gemeenschappelijke factor in de bifactor modellen klein waren en dat de gemeenschappelijke factor veruit de meeste variantie in het model verklaarde. Daarom concludeerden we dat de aanwezigheid van multidimensionaliteit niet interfereerde met een interpretatie van de BDI-II als een eendimensionale schaal. Sterker nog, de resultaten wezen er op dat het clusteren van items in meerdere dimensies en het gebruik van subschaalscores nauwelijks kan worden gerechtvaardigd. We adviseren psychologen daarom de totaalscore van de BDI-II te interpreteren als indicatie van depressiviteit, maar zorgvuldig om te gaan met interpretaties van subschaalscores.

In Hoofdstuk 5 onderzochten we de voorwaarden waaronder BDI-II verschilscores (voor- en nameting bij een behandeling) betrouwbare indicatoren kunnen zijn van individuele verandering in depressie in een steekproef van poliklinische patiënten. De resultaten laten zien dat voor de middelmatige en hoge depressiescores de meetprecisie hoog was, maar dat voor lage depressiescores de metingen erg onnauwkeurig waren. Ook vonden we dat verschillende ongewogen BDI-II scores kunnen wijzen op hetzelfde latente depressieniveau. Scores gebaseerd op ruwe somscores zijn bovendien niet geschikt voor intervalmeting. Ook vergeleken we de Z-test, een op IRT gebaseerde index voor het meten van statistisch betrouwbare verandering, met de vaak gebruikte reliable change index (RCI). Op

groepsniveau waren er geen grote verschillen tussen de RCI en de Z-Test, maar op individueel niveau kunnen de RCI en de Z-Test tot verschillende interpretaties van individuele verandering leiden.

Tot slot heb ik in de epiloog geprobeerd om de onderzoeksbevindingen uit deze studies te veralgemeniseren door de resultaten te vergelijken met andere studies. Ik beschrijf dat we door middel van moderne psychometrische methoden, zoals het GRM en het bifactor model, veel kunnen leren over de kwaliteit van vragenlijsten die we in de klinische praktijk dagelijks gebruiken om het psychisch functioneren van onze patienten te begrijpen en te volgen.

Dankwoord (Acknowledgements)

Voor mijn afstuderen aan de universiteit onderzocht ik het begrip hoop bij patienten in een dagklinische setting, maar al snel maakte ik mij veel te druk over de kwaliteit van de vragenlijst die we voor het onderzoek gebruikte. Inzichten vanuit moderne psychometrische theorieen waren nog nauwelijk gebruikt om te onderzoeken of klinische vragenlijsten kwalitatief goed in elkaar zaten. Met mijn IT achtergrond vond ik het leuk om met de (experimentele) software te werken waarmee deze moderne methoden toegepast kunnen worden. De evaluatie van vragenlijsten, die we in de klinische praktijk veel gebruiken, met behulp van moderne psychometrische methoden is zo het centrale thema geworden in mijn promotie onderzoek. Een proefschrift schrijven (als buitenpromovendus) is soms eenzaam werk, ik had het niet kunnen doen zonder de steun en hulp van velen. Ik wil jullie graag bedanken.

Allereerst, mijn promotor Rob Meijer. Rob, het is door jouw vertrouwen in mij en in ons onderzoek dat dit proefschrift er is. We zijn tegelijkertijd uit Twente vertrokken, jij naar Groningen en ik naar Amsterdam. Maar ondanks deze fysieke afstand was je altijd beschikbaar. Op mijn emails en concept verslagen die ik je kant op stuurde reageerde jij gedurende de gehele periode van mijn promotie altijd ontzettend snel. In onze (telefonische) afspraken gaf je me het gevoel dat er alle ruimte was voor overleg. Ik denk dat ons gemiddelde telefoongesprek ongeveer een uur duurde. Je weet me, in mijn enthousiasme en perfectionisme, enerzijds op geduldige wijze mijn gang te laten gaan maar anderzijds ook af te bakenen. Je hebt me veel bijgebracht over de psychometrie, het proces van wetenschappelijk onderzoek, schrijven en publiceren. Ik vond ook het leuk dat je me uitnodigde om gastcolleges te geven. Bedankt voor alles.

Als tweede wil ik Jolien Zevalkink bedanken. Je hebt me geholpen bij de aanlevering en het ordenen van de data die ik voor mijn onderzoek gebruikt heb. En ook jij hebt me geholpen om mijn werk af te bakenen. In het bijzonder heb je er als klinische wetenschapper voor gezorgd dat ik bleef nadenken over de praktische consequenties van onze bevindingen.

Ik wil ook Joost Baneke en Erwin Seydel bedanken. Joost, onder jouw begeleiding heb ik de eerste stappen gezet in het klinische werk. Ook heb jij mij gemotiveerd en het vertrouwen in me gehad om deze promotie te starten. Je hebt er persoonlijk garant voor gestaan dat dit gebeurde. Erwin Seydel, door jouw vertrouwen heb ik het project voort kunnen zetten.

Verder bedank ik graag de leden van de beoordelingscommissie, Peter de Jonge, Jan Henk Kamphuis en Klaas Sijtsma voor het aandachtig lezen van mijn proefschrift en jullie aanwezigheid tijdens mijn promotie. Ook bedank ik de overige leden van de promotiecommissie voor hun aanwezigheid.

Als buitenpromovendus ben ik verbonden geweest aan verschillende instanties. Aanvankelijk de Universiteit Twente en Mediant Geestelijke Gezondheidzorgs Oost en Midden Twente, later de Rijksuniversiteit Groningen en het Nederlands Psychoanalytisch Instituut. Ik ben er blij mee dat ik door deze samenwerkingsverbanden mijn onderzoek heb kunnen doen.

Ik wil Willem Heiser en de overige leden van het Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS) bedanken dat ik mocht spreken en luisteren op jullie halfjaarlijks congres. Door de gezelligheid en interessante presentaties van de IOPS promovendi heb ik me ingebed gevoeld in de psychometrische gemeenschap, dat heeft me zekerheid gegeven over mijn eigen onderzoek.

Caspar Berghout en Willemijn Hoek, bedankt voor jullie hulp bij het ordenen van mijn data. Anke Weekers, Iris Smits en Jorge Tendeiro, bedankt voor jullie hulp met de psychometrische analyses. Iris Egberink, Saskia de Maat, Esther en Caspar nogmaals, bedankt voor jullie hulp en tips bij het in elkaar zetten van mijn proefschrift. Bob, bedankt voor de prachtige omslag.

Mijn ouders wil ik bedanken voor alle inzet waarmee ze mij een veilige omgeving hebben geboden waarin ik mij heb kunnen ontwikkelen. Jullie hebben mijn leergierigheid gestimuleerd en belangrijke keuzes met mij gemaakt die ertoe geleid hebben dat ik hier vandaag sta. Verder wil ik alle familie, vrienden en collega's bedanken die er over de jaren met hun interesse en steun voor mij zijn geweest. In het bijzonder Frans, Rolf en Esther; bedankt voor jullie continue aanwezigheid.

"The Clinician's passion for complexity is almost certainly a valid way to conceive of the universe. The Psychometrikan's passion for reduction is a practical compromise, to simplify problems enough so that scientific methods can come to grips with them."

(Cronbach,1954, p. 266)